Interpretable Machine Learning in Credit Risk Modelling

Giulio Bellini

MSc Financial Technology

University of Birmingham

Table of Contents

Table of Figures

Interpretable Machine Learning in Credit Risk Modelling

Introduction

Machine learning has marked a turning point in the development of artificial intelligence by empirically approaching tasks that were considered almost impossible to perform procedurally (Mullainathan & Spiess, 2017). Intelligent tasks such as visual and audio recognition, translation, text and image generation are performed by machines not according to rules deduced by humans, but based on a function inductively estimated from large amounts of data. For example, it is possible to train a machine learning model to classify images as containing a cat or a dog by applying an algorithm that estimates a function $f(x)$ that relates the pixels $x_1, x_2, \dots, x_n$ of an image to the presence of a cat or a dog y. The algorithm is applied to a large dataset of labelled images and the function $f(x)$ is estimated empirically. This approach makes machine learning similar to econometric (Mullainathan & Spiess, 2017). This has led to the use of machine learning models in economics and finance, where the tasks and nature of the data are different, and where historically econometric models such as linear and logistic regression have been effective. Firstly, machine learning (at least supervised machine learning) is concerned with the task of prediction (Mullainathan & Spiess, 2017): predicting a value or label y from a set of inputs $x_1, x_2, \dots, x_n$. It does this by capturing generalisable patterns in the data by fitting complex and flexible functions. Many problems in economics, however, involve estimating parameters (Mullainathan & Spiess, 2017), such as the parameter β that expresses the relationship between the independent variable x and the dependent variable y. Econometrics therefore focuses on making inferences about causal relationships between variables. Machine learning models are not built to discover causal relationships, but to exploit complex correlations to generate high quality predictions. The parameters of a machine learning model should therefore not be interpreted as representative of the underlying data-generating process, as they are based on strong assumptions that make them unstable (Mullainathan & Spiess, 2017). However, they can be used to interpret the prediction function and the patterns it captures. This is particularly important when applying machine learning to high-impact prediction tasks in domains such as medicine, law and business. In these fields, interpretability, coupled with causal inference methods, allows the identification of biases in the data and the model, and increases confidence in the use of machine learning for informed decision making.

We can therefore conclude that machine learning should be applied to problems where quality predictions are of great value (Mullainathan & Spiess, 2017). One such problem is the prediction of borrower default. A lender's ability to distinguish between creditworthy and risky applicants, or to estimate the future default rate of its loan portfolio, is of great financial value. Historically, probability of default (PD) has been estimated using statistical models, the most common of which is logistic regression, which models the log odds of an event as a linear combination of independent variables. However, there is growing interest in applying machine learning to this task. Models such as random forests and gradient boosting machines, which are ensembles of simple decision

trees, are state of the art for classification tasks involving tabular data (Grinsztajn, Oyallon, & Varoquaux, 2022), such as those typically used in credit risk. Tree-based models have the advantage of capturing complex non-linear relationships and interactions between variables, which can improve their predictive performance over linear models such as logistic regression. The recent academic literature is abundant with examples of the application of machine learning to credit risk. The following is a non-exhaustive review of the literature.

(Khandani, Adlar, & Andrew, 2010) applies the Classification and Regression Tree (CART) algorithm to the prediction of credit card account defaults and estimates the cost savings of using the model in the decision to cut credit lines. (Barbaglia, Manzan, & Tosetti, 2023) show how Gradient Tree Boosting and Extreme Gradient Boosting outperform logistic regression in predicting the default of European residential mortgages. (Sirignano, Sadhwani, & Giesecke, 2020) explore the use of neural networks to capture highly non-linear relationships in a large dataset of US mortgages and predict their default. (Ojha & JeongHoe, 2021) compare several conventional machine learning models (logistic regression, random forest, linear discriminant analysis, k-nearest neighbours) and deep learning models (multilayer perceptron, convolutional neural network, recurrent neural network, long short-term memory) as well as an ensemble of them for predicting default of US residential mortgage-backed securities, showing that random forest, multilayer perceptron and ensemble perform best, although the gap between the different models is small. (Butaru, et al., 2016) build PD models for credit cards across six major commercial banks, showing how C4.5 algorithm decision tree and random forest outperform logistic regression, but that a single model cannot be effectively applied to all institutions as risk factors and predictability vary. Finally, (Lessmann, Baesens, Seow, & Lyn C., 2015) presents probably the most comprehensive comparison of classification algorithms for PD modelling. The paper extends previous work (Baesens, et al., 2003) and the subsequent extensive literature on retail credit scoring by including novel classifiers (such as homogeneous and heterogeneous ensembles), using multiple datasets of considerable size, and considering conceptually different performance metrics. A total of 41 algorithms, totalling 1141 models (the same algorithm can produce different models depending on the set of hyperparameters), are benchmarked on eight retail credit scoring datasets. Algorithms are ranked across datasets and performance metrics. The best overall classifier is HCES-Bag (Caruana, Munson, & Niculescu-Mizil, 2006), a direct selective ensemble, while random forest and artificial neural network are ranked as the best homogeneous ensemble and best individual classifier respectively. The results from the literature thus suggest that machine learning models are superior to traditional ones in modelling the probability of default.

However, machine learning models have a drawback. They are often too complex to interpret and are thus referred to as "black box" models. The inability to understand the inner workings of a model not only makes it difficult to troubleshoot during the development and validation phases, but also to ensure regulatory compliance. The credit sector is highly regulated and involves decisions with high economic and social impact. For this reason, it is important to understand how a model generates predictions from the independent variables. Based on this need, several works

have applied Explainable AI methods to explain the predictions of black box models. (Misheva, Hirsa, Osterrieder, Kulkarni, & Lin, 2021) use Local Interpretable Model Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) and Shapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) to generate local and global explanations of XGBoost (Chen & Guestrin, 2016) and neural network predictions of US consumer loan default. (Babaei, Giudici, & Raffinetti, 2023) use the global SHAP values of the independent variables to perform a stepwise variable selection algorithm. In this way, they develop a sparse random forest to predict the default of European SMEs without sacrificing too much performance and outperforming the baseline complete logistic regression model. (Bussman, Giudici, Marinelli, & Papenbrock, 2021) use the Shapley values to group the model's predictions according to the similarity of their explanations. They develop a XGBoost model to score SMEs based on credit risk and group the predictions based on the Euclidean distance between the vectors of variable contributions.

An alternative approach to the interpretability problem in machine learning is to create inherently interpretable models. The focus shifts from building complex models to building complex algorithms that produce interpretable models with performance comparable to black box models. For example, (Chen, et al., 2018) develop a transparent "two-layer additive risk model" that is decomposable into meaningful subscale models and provides three types of explanations consistent with the global model. With this fully interpretable global model, the authors won the FICO Recognition Award in the 2018 FICO Explainable Machine Learning Challenge[1].

This paper draws on the literature on interpretable machine learning and focuses on the development of an inherently interpretable model for predicting the probability of default of a portfolio of mortgages. The main contribution of this work is the use of a large real-world dataset of mortgage origination and performance data, rather than an experimental dataset. The performance of the interpretable model has been compared with that of state-of-the-art black box models, showing how increasing interpretability positively affects performance. The following sections of the paper provide the theoretical background for interpretable machine learning and the development of credit risk scorecards. This is followed by a description of the data collection process and the creation of development samples. Next, the interpretable model and the two black box models are presented. Finally, the performance of the models is compared and the structure of the interpretable model is examined.

[1] FICO Explainable Machine Learning Challenge overview https://community.fico.com/s/explainable-machine-learning-challenge

Interpretable Machine Learning

Definition of interpretability

Machine learning is being used in a wide range of fields, from hard sciences such as physics and biology to soft sciences such as business and economics. It has proven to be extremely effective at prediction tasks, being able to learn complex relationships from large amounts of structured and unstructured data. More recently, however, practitioners have realised that in addition to making accurate predictions, machine learning models can provide valuable insights into the underlying relationships in the data, called "interpretations" (Murdoch, 2019). Interpretations allow one to understand what the machine learning model has learned from the data and are particularly important in areas where high-stakes decisions are made (Rudin, 2019), such as medicine, criminal justice and lending, or where it is necessary to audit the models and their predictions to ensure compliance with regulations. In these cases, interpretations allow the model to be troubleshooted with the support of domain experts and build confidence in its predictions (Murdoch, 2019). In medicine, for example, it would be essential to be able to interpret a model predicting cancer risk based on a patient's clinical records or x-rays, otherwise one would have to blindly trust the model with the patient's life. Similarly, in the lending sector, a model used to make credit decisions (e.g. whether to grant a loan) must be interpretable so that it can be audited to ensure that it does not discriminate on the basis of gender or ethnicity. Despite the growing interest, there is still no clear definition of interpretability in machine learning. The field of interpretable machine learning is attributed to a wide range of methods to create both interpretable models and interpretations of their predictions. (Murdoch, 2019) defines interpretable machine learning as:

"The extraction of relevant knowledge from a machine-learning model concerning relationships either contained in the data or learned by the model."

Knowledge is considered relevant if it provides insight into a particular problem to a particular audience who can use it to make informed decisions, communicate more effectively or make new discoveries (Murdoch, 2019). It is therefore clear that interpretations are a fundamental part of the data science lifecycle, i.e. the process of knowledge extraction. However, this definition is too broad. In fact, interpretable machine learning focuses on what (Murdoch, 2019) calls "model-based interpretability", which focuses on creating inherently interpretable models (e.g. sparse linear models or sparse decision trees) that provide a complete and immediate view of the learned relationships. "Post-hoc" interpretability concerns the interpretation of black-box models (non-interpretable, such as neural networks or ensembles) and their predictions using methods based on function approximation and derivatives (Rudin, et al., 2022). Methods for post-hoc interpretability are grouped under the term "Explainable AI" (XAI). While Interpretable Machine Learning and Explainable AI are often confused or lumped together, they represent two distinct approaches to creating interpretations. They use different methods at different stages of the data science lifecycle, and each has its pros and cons, as we'll see below. However, they share the goal of providing an understanding of the relationships learned by the model (descriptive accuracy) without

compromising its performance (predictive accuracy) (Murdoch, 2019). There is a myth about the existence of a trade-off between predictive and descriptive accuracy (interpretability-accuracy trade-off), motivated by the idea that black-box models tend to have higher predictive accuracy than fully interpretable models. However, this trade-off does not seem to exist in practice (Rudin, et al., 2022). On the contrary, the transparency of interpretable models favours the discovery of useful relationships in the data or bugs in the model that can be used to optimise its predictive performance (Rudin, et al., 2022). A full discussion of the trade-off between interpretability and accuracy can be found in the dedicated chapter.

This paper focuses on interpretable machine learning rather than Explainable AI. (Rudin, et al., 2022) provides a more specific definition of *Interpretable Machine Learning*:

"An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain."

In mathematical terms, an interpretable supervised machine learning model f solves the following minimisation problem:

$$min_{f \in F} \frac{1}{N} \sum_{i} Loss(f, x_i, y_i) + C * InterpretabilityPenalty(f)$$

$$subject\ to \quad InterpretabilityConstraints(f)$$

where the loss function, the interpretability penalty and the interpretability constraints are domain specific. Consequently, we can define a model as "uninterpretable" or "black box" if its formula is too complicated for the human mind to understand or is proprietary and therefore inaccessible (Rudin, et al., 2022). The inner workings and predictions of a black box model are inherently opaque, uninterpretable. There is no single definition of the interpretability of a machine learning model that applies across all domains. The definition varies by domain, and within the same domain it may vary by the use case of the model. It is up to the model developer, with input from domain experts, to choose the most appropriate definition and interpretability metrics, just as they choose the most appropriate predictive performance metrics (e.g. accuracy, AUC, F-score, mean absolute error, etc.) for the domain and prediction task at hand (Rudin, et al., 2022). Measures of interpretability include constraints such as model sparsity, monotonicity of predictions with respect to one or more variables, and decomposability into sub-models (Rudin, et al., 2022). For example, in the credit domain, a lender seeking to develop an interpretable model to rank the loans in its portfolio according to their probability of default might impose constraints on the number of variables used (sparsity) so that the predictions can be monitored by a credit expert, monotonicity constraints on variables considered to have high predictive power (e.g. the predicted probability of default must be directly proportional to the loan-to-value (LTV) ratio and inversely proportional to the borrower's credit score), or constraints on the exclusion of protected attributes (e.g. ethnicity, gender, religion) and their proxies (e.g. zip code) from the model. Credit risk often involves

modelling tabular/structured data, i.e. data organised in tables with meaningful real or discrete features (e.g. LTV, DTI, credit score, loan balance, interest rate), rather than raw/unstructured data such as text or images where individual features (e.g. words, pixels) are not meaningful in themselves. For tabular data with meaningful features, the main measure of interpretability is sparsity (Rudin, et al., 2022). A model is defined as "sparse" if only a small fraction of its parameters is non-zero (Sharan, Tai, Bailis, & Valiant, 2017). The features associated with these parameters can therefore be interpreted as the most meaningful (Murdoch, 2019). Sparsity allows one to understand how variables contribute together rather than individually (Rudin, 2019). Sparsity can be achieved by adding a penalty term to the loss function, as in the case of LASSO regression, or by using model selection metrics that favour sparse models, such as the Akaike Information Criterion (AIC) or adjusted R-squared (Murdoch, 2019). Sparsity constraints are particularly useful in problems involving high-dimensional data (large number of features) (Murdoch, 2019), as they allow the selection of a small subset of important features from the original feature space, making the model simpler, more interpretable and often more robust.

The chosen functional form of the model is also an example of an interpretability constraint (Rudin, et al., 2022). For many problems involving tabular data, it is possible to create fully interpretable models, such as sparse linear models or sparse decision trees, which perform as well as black-box models such as random forests and neural networks, as discussed in more detail below. The functional form can introduce two sources of interpretability into the model: simulatability and modularity (Murdoch, 2019). A model is simulatable if its decision-making process (producing a prediction from an input) can be simulated internally by the human mind (Murdoch, 2019). A simulatable model must have a small number of parameters that a human needs to keep track of. For example, a sparse decision tree with low depth can be simulated because its decision process consists of classifying an input based on logical conditions built on a few attributes (e.g. an applicant is classified as high risk by default because he has LTV > 80%, Credit Score < 600 and DTI > 50%). A model is said to be modular if parts of its decision process can be interpreted independently (Murdoch, 2019). General linear and additive models are modular models because they generate predictions by adding the contribution of each individual variable (Murdoch, 2019). The influence of an individual variable on a prediction can therefore be interpreted independently.

Sparsity, monotonicity, simulatability and modularity are just some of the constraints that can be used to increase the interpretability of a machine learning model in a given domain. This paper follows these principles to create an interpretable machine learning model to discriminate mortgages based on their risk of default.

Importance of Interpretability

Interpretable machine learning is crucial in domains involving high-stakes decisions and, more generally, in model troubleshooting (Rudin, et al., 2022). It is possible to distinguish the application domains of machine learning according to the nature of the problems they seek to solve. Some domains involve low-stakes decisions (e.g. advertising, e-commerce, entertainment)

whose economic and personal consequences are limited. These domains were the first to use machine learning models to create accurate prediction systems for recommending products, setting the target of an advertising campaign, etc. (Rudin & Radin, 2019). Other domains, however, involve high-stakes decisions (e.g. healthcare, finance, criminal justice) that can have a significant impact on people's lives (Rudin, 2019). For example, the decision to treat or not to treat a patient based on a machine learning model has a direct impact on the patient's health. In finance, wrong decisions can lead to huge losses, and in criminal justice to wrongful convictions. In these domains, the use of interpretable models is crucial. In finance, for example, econometric models such as linear and logistic regression have traditionally been used, which are interpretable and focus on capturing causal relationships between variables. More recently, however, there has been a shift in these fields towards the use of black-box machine learning models. Based on the belief that black-box models consistently outperform interpretable models, deep neural networks have been used to predict cancer risk from X-rays and complex tree ensembles to predict credit decisions.

The main problem with black-box models, which makes interpretable models crucial for high-stakes decisions, is the difficulty of debugging them. If the inner workings of a model cannot be understood, it is very difficult, if not impossible, to confirm that the model has learned real relationships in the data, or only spurious patterns and biases specific to the dataset on which it was trained (Rudin, et al., 2022). It is therefore difficult to determine whether the model will perform equally well on new, unseen data (generalisation ability) or whether it will be robust to changes in the underlying distribution of the data (domain shift) when used in practice (Rudin, et al., 2022). It is also difficult to confirm that the model does not use restricted variables, or proxies for them, to ensure the fairness and regulatory compliance of the predictions (Rudin, et al., 2022). For example, a model should not use information about an individual's gender, ethnicity or socio-cultural background to make lending or criminal justice decisions. The use of a black-box model therefore imposes trust in its predictions despite having limited information about the process by which they are produced (Rudin, et al., 2022). In contrast, an interpretable model can be easily debugged because its decision-making process is transparent.

The ability to debug a model becomes even more important when considering the entire data science lifecycle (Rudin, et al., 2022), displayed in *Figure 1*. The data science lifecycle can be broadly defined as the iterative process of extracting knowledge from data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). It is the process used in practice to solve real problems with data and machine learning. It starts with the definition of a domain-specific problem (e.g. accurately predicting the probability of a borrower defaulting) (Murdoch, 2019). The next step is to create the dataset needed to build the model. This phase includes data collection, ensuring representativeness with respect to the population of interest, and pre-processing (data cleaning, standardisation, transformation, etc.) of the data. This is followed by model development, in which the functional form is chosen, the relevant features are selected, and the model is fitted. Finally, there is the model evaluation phase, in which the patterns found are analysed, the accuracy of the predictions is assessed, and the model is debugged. Interpretability plays a key role in the model development

and evaluation phases (Murdoch, 2019). The data science lifecycle is iterative because the model may not be able to solve the domain problem on the first interaction (Murdoch, 2019). Furthermore, data analysis and interpretation of the model and its predictions may reveal errors in the process. In both cases, the process needs to be updated and repeated. Choosing a black box model can hinder the iterative process by making the interpretation phase difficult, leading to the development of a suboptimal model. Using an interpretable model, on the other hand, allows for troubleshooting, which leads to greater accuracy and robustness of the model, and thus higher quality decision making (Rudin, et al., 2022).

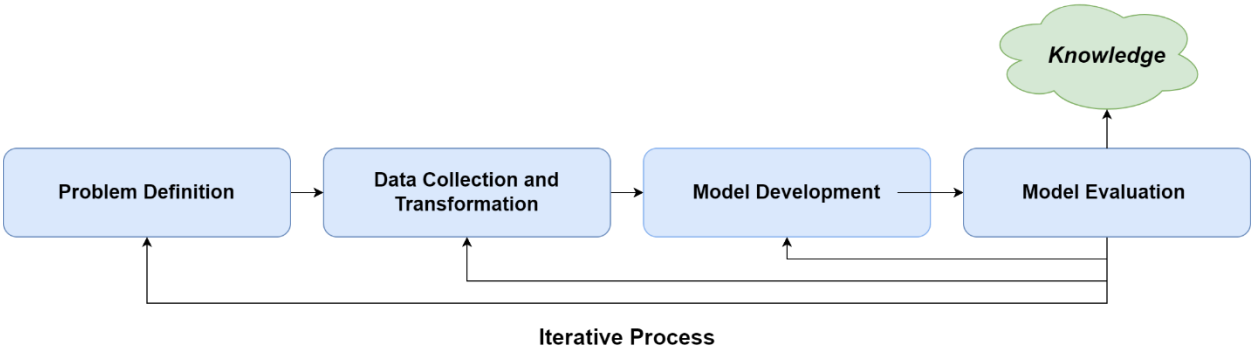

Figure 1: Data Science Lifecycle

Interpretable ML and Explainable AI

The difficulty of understanding and debugging black-box models has recently led to the emergence of "Explainable AI" (XAI) (Rudin, 2019). XAI does not involve the creation of inherently interpretable models but focuses on explaining a black-box model using post-hoc model approximations (local or global), derivatives, feature importance measures, and other statistics (Rudin, et al., 2022). Explainable AI methods often consist of locally approximating a black box model (e.g. LIME) or globally approximating its predictions (e.g. SHAP values) using a separate, simpler, more interpretable post-hoc model (e.g. linear model). Obviously, the model used to explain the black box cannot generate explanations that are 100% accurate, since in that case the simple post-hoc model would be the same as the original complex model and could be used in its place (Rudin, 2019). Explanations therefore never perfectly reflect the inner workings of the black box model, and sometimes even use a different set of features (Rudin, 2019). When trying to explain a black box model, one has access to its input (data, features) and output (predictions). The formula that generates predictions from the data by combining the features is too complex for the human mind to understand or is proprietary (Rudin & Radin, 2019). Therefore, rather than explaining the actual calculations of the original model, Explainable AI methods only show relationships between predictions and features. These relationships may be very different from those used by the model to make predictions (Rudin, 2019). The inaccuracy of the explanations creates uncertainty and mistrust because it is difficult to distinguish cases where the model's predictions are correct and the explanations are incorrect from cases where the predictions are

incorrect and the explanations are correct (Rudin, 2019). Using an inaccurate explanation model therefore reduces, rather than enhances, confidence in both the explanations and, by extension, the black box model they are intended to explain (Rudin, 2019). It also adds a layer of complexity to the debugging process, since having inaccurate explanations forces one to debug both the black box model and the explanation model (Rudin, 2019).

The field of Explainable AI arose as an attempt to address the non-interpretability problems of black-box models in order to justify their use in high-stakes decision domains, rather than focusing on developing interpretable models (Rudin, 2019; Rudin & Radin, 2019). The emergence of the XAI reflects the tendency of practitioners and academics to use black box models even when they are not needed. There are several reasons why people tend to prefer black box models, but the main one is the widespread belief that more complex models have better predictive performance. However, this is often not the case, especially when modelling structured data (Rudin, et al., 2022).

Accuracy-Interpretability Trade-Off

The success of deep learning in complex tasks involving unstructured data, such as computer vision and natural language processing, has led to a widespread perception that complex and uninterpretable machine learning models are necessary to maximise prediction accuracy (Rudin, 2019; Murdoch, 2019; Rudin & Radin, 2019). This has led to the myth of the existence of a trade-off between accuracy and interpretability in machine learning models, which is not supported by any scientific evidence (Rudin, et al., 2022). This is particularly true when considering the full data science lifecycle, i.e. the development of machine learning to solve real-world problems. In these cases, interpretability improves accuracy, not degrades it, by enabling troubleshooting and subsequent iterative process updates (Rudin, et al., 2022). Machine learning models, on the other hand, are typically benchmarked on static datasets that are not iteratively updated but are used to compare algorithms in an experimental environment (Rudin, et al., 2022). However, even in these cases, the use of interpretable models does not appear to lead to a deterioration in predictive performance, particularly when modelling tabular data. Complex black-box and simple interpretable models tend to perform similarly when the data is structured and has meaningful features (Rudin, 2019), as in the case of credit data (e.g. credit score, LTV, DTI, interest rate). The myth of the trade-off between accuracy and interpretability arose mainly from the association of interpretability with sparsity (Rudin, et al., 2022). In fact, sparsity is a measure of the simplicity of a model, and there is almost always a trade-off between accuracy and sparsity (Rudin, et al., 2022). Sparsity, however, is only one of several measures of interpretability, the definition of which is broader and domain dependent. An interpretable model may be sparse but at the same time have other properties (e.g. monotonicity, decomposability, adherence to business logic) that may increase its accuracy and make it preferable to black box models. The false dichotomy between accuracy and interpretability has led researchers and practitioners not to even attempt to create interpretable models for complex problems, but to see black box models as the only alternative (Rudin, et al., 2022). This allows companies to sell complex proprietary models for high-stakes

decisions (e.g. the COMPAS model for predicting recidivism[2]) when very simple interpretable models with the same performance exist (Rudin & Radin, 2019; Angelino, Larus-Stone, Alabi, Margor, & Rudin, 2018), with potentially harmful consequences for people's lives. There is also a trend in credit scoring away from traditional econometric methods, which are inherently interpretable and designed to uncover causal relationships, in favour of black-box machine learning models such as ensembles and neural networks. Explanatory AI methods such as LIME and SHAP scores are often coupled with these models in an attempt to explain their inner workings or predictions.

The Preference for Black-Boxes

In addition to the alleged trade-off between accuracy and interpretability, there are other reasons why black-box models flanked by explanatory models are so often preferred to interpretable models, as described in (Rudin, 2019). First, companies can profit from selling proprietary models and therefore have no interest in stimulating the search for alternative interpretable and transparent models. Proprietary models are also used for high-stakes decisions, as in the case of COMPAS, without the producing companies necessarily being accountable for the quality of their predictions (Rudin, 2019). Black box models are also seen as more resistant to manipulation. In reality, what prevents a model from being manipulated is its ability to capture real patterns in the data, not the fact that it is incomprehensible (Rudin, 2019). For example, a credit risk model that predicts that the risk of default on a loan decreases as the credit score increases or the DTI decreases will encourage applicants to behave in ways that increase their creditworthiness, not just their chances of getting the loan. Finally, black-box models often require less effort and time to develop (Rudin, et al., 2022). By definition, interpretable models respect domain-specific constraints. Building an interpretable model therefore requires solving constrained optimisation problems, which are often computationally harder than unconstrained problems (Rudin, 2019). In addition, constructing the interpretability definition for the domain and debugging the data and model takes time (Rudin, et al., 2022). It is often easier to use black box models than to debug and solve hard constrained optimisation problems (Rudin, 2019). However, in high-stakes decision domains such as credit, the cost of human and computational resources required to develop interpretable models is less than the cost of implementing a flawed model (Rudin, 2019).

[2] Correctional Offender Management Profiling for Alternative Sanctions tool (COMPAS) documentation available at https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx

Credit Risk Scorecards

Credit scorecards are mathematical models that separate individuals based on their creditworthiness (Siddiqi, 2017). Scorecards are developed from a set of individual characteristics that are statistically significant in separating riskier from less risky accounts (Siddiqi, 2017). Such characteristics may include demographic information (e.g., zip code, income, employment status), information about current or past financial behaviour (e.g., payment history, delinquency status, total debt, debt-to-income ratio), credit bureau data (e.g., credit score), real estate data (e.g., property value, location), and potentially any other data source to which the financial institution creating the model has access (Siddiqi, 2017). Attributes are then extracted from the characteristics (e.g. Loan-To-Value (LTV) > 80% is an attribute of the LTV characteristic) to which a score is assigned, i.e. a value obtained through statistical analysis or a training algorithm in the case of machine learning based models (Siddiqi, 2017). An individual's total score is the sum of the scores for each of its attributes. The higher an individual's score, the higher their risk.

Based on a borrower's score, the lender can implement strategies to minimise the risk of default and thus potential losses (Siddiqi, 2017). For example, a lender may refuse to lend to a new applicant or offer payment deferral solutions to its existing creditors to reduce the risk of insolvency if they have a score above a certain threshold. Although the availability of a scorecard allows for the full automation of processes such as the decision to approve or deny a loan, in many cases, particularly for high-value credit products such as mortgages, organisations tend to use the score as one of several measures to assess the creditworthiness of applicants, with significant human involvement and judgement (Siddiqi, 2017).

There are two main types of statistical models used in the development of credit scorecards: classification models and duration models. Classification models aim to accurately distinguish high default risk accounts from low-risk accounts. Classification models include logistic regression, which can be considered the industry standard in credit scorecard development. Duration models, on the other hand, aim to predict when an account will default. The most popular approach to building duration models is survival analysis, which includes models such as the proportional hazards model that can predict the time to default. This paper focuses exclusively on the development and comparison of classification models, in particular Generalised Additive Model (GAM), Random Forest (RF) and Extreme Gradient Boosting Machine (XGBoost).

According to (Siddiqi, 2017), scorecards are the most used credit risk modelling method in the industry due to their ease of interpretation and implementation. This is because credit scorecards are traditionally developed using statistical modelling techniques, such as logistic regression and decision trees, which are inherently interpretable. Their predictions can be explained to stakeholders (management, model validators and other staff), regulators, auditors and customers, and their development and validation processes are well understood and not a black box (Siddiqi, 2017). Traditional credit scorecards are therefore both interpretable and explainable. In recent years, however, new types of models from the field of machine learning have become increasingly

popular in credit risk modelling, namely tree-based ensemble models and neural networks. The explosion in popularity of these models can be explained by their remarkable accuracy in classification tasks involving both structured and unstructured data (Grinsztajn, Oyallon, & Varoquaux, 2022). This has led to widespread industry interest in applying these methods to the development of credit scorecards. However, such models are often too complex for the human mind to understand. They are therefore uninterpretable black-box models that make the iterative process of model development and validation extremely difficult, if not impossible. Moreover, black-box models are hardly compliant with the strict regulation that characterises the credit sector, especially retail sector (Siddiqi, 2017). For example, it is extremely complex to audit a black-box credit model developed from potentially thousands of attributes to ensure that it does not use protected attributes such as gender and ethnicity, or proxies of them, in making predictions (Fuster et al., n.d.).

Generic and Custom Scorecards

When a financial institution needs a credit risk model, for example to streamline its lending operations or for risk management or regulatory compliance, it has two alternatives. The first is to buy the model from an external credit risk vendor that develops generalist models using large samples of credit bureau data drawn from a population similar to the financial institution's customer base (Glennon, 2008). The second option is to develop the model 'in-house'. The development of in-house credit risk models often involves significant development costs (Siddiqi, 2017), mainly related to the hiring of specialised staff, but has several advantages. The main advantage is the ability to develop the model on a sample selected exclusively from the financial institution's own historical customer data, often combined with data from credit bureaus (Glennon, 2008). In this way, the institution can create a tailor-made model to predict the behaviour of its current and future clients.

In recent years, there has been a trend towards the development of in-house models, not only by large financial conglomerates, but also by small and medium-sized institutions (Siddiqi, 2017). This trend can be explained, at least in part, by regulatory developments. In 2004, the Bank for International Settlements published the Basel II Accord[3], which established a new framework for determining the minimum amount of capital that banks must hold to cover the risks associated with their lending and investment activities in order to minimise the risk of insolvency. This was followed by Basel III[4], published in 2010, which increased minimum capital requirements in the wake of the 2008 financial crisis. Since the publication of Basel II, many banks around the world have begun to comply, either mandatorily or voluntarily, with the Foundation (F-IRB) or Advanced (A-IRB) Internal Ratings Based approaches defined by the Accord (Siddiqi, 2017). Under the F-IRB, banks may develop internal estimates of probability of default (PD) for individual or pools of credits, while under the A-IRB they may also model loss given default (LGD) and exposure at

[3] Full text of Basel II: Revised international capital framework available at https://www.bis.org/publ/bcbsca.htm
[4] Full text of Basel III: international regulatory framework for banks available at https://www.bis.org/bcbs/basel3.htm

default (EAD). Financial institutions that have chosen to comply with the Internal Ratings Based approaches of Basel II have moved from purchasing credit scorecards from credit risk vendors to developing them in-house (Siddiqi, 2017). This allows banks to leverage their knowledge of internal data and business acumen to create more accurate models. The scorecard development process can also provide unique business insights that add to the organisation's knowledge base (Siddiqi, 2017).

Due to data availability constraints, this paper develops generic credit risk models. However, the aim is to demonstrate how model interpretability facilitates the process of knowledge discovery without degrading performance, in order to encourage financial institutions to develop internally interpretable credit scorecards without indulging in the rhetoric of the interpretability-accuracy trade-off.

Interpretability Importance in Credit Risk

A credit risk model is a business tool used for better decision making. Its workings need to be understood by both developers and decision makers to ensure regulatory compliance and operation in line with business constraints and logic, and to enable debugging by development and validation teams. In this regard, (Siddiqi, 2017) states:

> *" ... since the credit crisis of 2007-2008, the tolerance at most banks for complex/black box models and processes is gone. The business user expects a model that can be understood, justified, and where necessary, be tweaked based on business considerations..."*

It is therefore important that credit scorecards are inherently interpretable and not complex and incomprehensible mathematical models developed in isolation by a small group of experts.

The credit sector is highly regulated and involves high-stakes decisions. Granting a loan to a customer who is misclassified as creditworthy can result in huge losses for a bank, just as denying a loan to a customer who is misclassified as uncreditworthy can have a major negative impact on his or her life. The importance of developing accurate yet interpretable credit risk models to make accurate and explainable predictions is therefore clear.

The Financial Stability Board (FSB), the international body that oversees the global financial system, examined the potential financial stability implications of the use of AI and machine learning in financial services (FSB, 2017). While recognising their positive effects, such as contributing to the efficiency of the financial system, it expressed concern that "the lack of interpretability or verifiability of AI and machine learning methods could become a macro-level risk". Focusing specifically on credit scoring applications, the FSB notes that AI has made it possible to incorporate alternative sources of data, often unstructured, into credit assessments. This has helped to increase access to credit for individuals with limited credit histories, who have traditionally been considered unscorable. At the same time, however, it has raised concerns about privacy and the introduction of bias into credit decisions. The FSB also points out that black box machine learning models have not been proven to outperform traditional methods.

There are also laws that require a certain level of model interpretability. In the US, the Equal Credit Opportunity Act[5] not only prohibits lenders from discriminating against applicants on the basis of protected characteristics (gender, race, age, religion, etc.), but also establishes the right to know the reason for denial, requiring the lender to provide the applicant with a specific reason for rejection. In these cases, the use of black box models together with explainable AI methods may not be sufficient and an interpretable model would be necessary to comply with the regulations (Croxson, Bracke, & Jung, 2019).

Application and Behavioural Scorecards

Credit scorecards can be divided into two categories according to the purpose for which they are created. The first type is application scorecards (Siddiqi, 2017), which are used when a credit application is made to estimate the applicant's likelihood of default over the lifetime of the loan. On the basis of the applicant's score, the lender can decide whether or not to grant the loan, the interest rate to be offered or the amount of the advance to be requested. Application scorecards are developed based on demographic and financial data available at the time of application and sampled from past applications. The second type is behavioural scorecards (Siddiqi, 2017), which are used after the loan is granted and throughout its life to estimate the borrower's likelihood of default over a period of time, called the outcome window (Kennedy, Mac Name, Delany, O'Sullivan, & Watson, 2013). Using behavioural scorecards, the lender can make risk management decisions for the specific loan or account, such as offering the borrower a payment deferral, to reduce the likelihood of default and minimise associated losses. Behavioural scorecards are developed using historical loan repayment performance data observed at a point in time and related to a past time window, called a performance window (Kennedy, Mac Name, Delany, O'Sullivan, & Watson, 2013). A quantitative measure of creditworthiness is used in the development of both application and behavioural scorecards. For example, we can define a borrower as "defaulted" if they miss more than three consecutive payments in the 12 months following the application/observation date and create a variable “Default” that takes the value 1 in the case of default and 0 otherwise.

This work focuses on the development of behavioural scoring models, developing and comparing interpretable and black-box models from a large original dataset of mortgage data, including both application and performance data, as well as local economic data.

[5] Full text of Equal Credit Opportunity Act available at https://www.justice.gov/crt/equal-credit-opportunity-act-3

Dataset Construction

Data Source

Scorecards are developed on the assumption that future performance is reflected in past performance, so the historical performance of loans made in the past is analysed to predict the performance of current and future loans in the future (Siddiqi, 2017). In developing a generic behavioural scorecard, it is therefore necessary to collect data on a portfolio of loans (mortgages) that is as similar as possible to that of a hypothetical mortgage issuer or investor. The main data source for this work is Freddie Mac's Single-Family Loan-Level Dataset[6], a large public dataset of mortgages that Freddie Mac has purchased from lenders across the United States since 1999. It is a large and comprehensive sample for the development of generic scorecards that can be used by US issuers and investors.

Freddie Mac[7] (short for Federal Home Loan Mortgage Corporation) is a US government-sponsored enterprise (GSE), along with Fannie Mae[8] (Federal National Mortgage Association). It was created in 1970 to expand the secondary market for US mortgages and has been under conservatorship of the Federal Housing Finance Agency (FHFA) since 2008. Both GSEs have the same mission of providing liquidity and stability to the mortgage market. As described by the FHFA[9], they "purchase mortgages from lenders and either hold these mortgages in their portfolios or package the loans into mortgage-backed securities (MBS) that can be sold to investors". In this way, lenders can use the liquidity generated by the sale to make additional loans and expand their businesses, ensuring a stable and affordable supply of mortgages for individuals and families (FHFA, 2024). The GSEs assume the credit risk of the mortgages they purchase by guaranteeing the timely payment of principal and interest to the MBS holders, who are thus exposed only to market risk (e.g. market devaluation of the loans due to an increase in interest rates). Hedging against credit risk has the effect of attracting new investors to the secondary mortgage market by making it more liquid, increasing the funds available to the housing market and helping to reduce the interest paid by borrowers (FHFA, 2024). (Bartlett, 2021) details the process of selling mortgages to the GSEs and underwriting credit risk. The process begins with the lender submitting the applicant's information (credit score, LTV, debt-to-income, loan amount, etc.) to one of the GSEs' automated underwriting systems (Loan Prospector for Freddie Mac). If the underwriting system approves the application, the lender can make an offer to the applicant. If the applicant accepts the offer and the mortgage is originated, the lender immediately sells it to the GSE and receives a lump sum of cash (cash window acquisition) or, if the lender sells large volumes of loans in bulk, an MBS comprised primarily of the loans themselves (MBS swap acquisition channel) (FHFA, 2024). An important part of the offer made to the applicant is the interest rate. In

[6] Data and Documentation for the Freddie Mac's Single-Family Loan-Level Dataset available at https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset
[7] https://www.freddiemac.com
[8] https://www.fanniemae.com
[9] FHFA: About Freddie Mac & Fannie Mae, available at https://www.fhfa.gov/about-fannie-mae-freddie-mac

the GSE mortgage market, the interest rate is the sum of three components: (1) the base mortgage rate for mortgages guaranteed by the GSEs (the credit-risk-free rate), (2) a guarantee fee (g-fee) paid by the lender to the GSEs to cover operational costs and potential credit losses related to borrower default, and (3) a discretionary rate component charged by the lender to cover its own costs or for strategic reasons (Bartlett, 2021). The g-fees depend on the product type (e.g. 30-year fully amortising fixed-rate mortgage), loan-specific characteristics, in particular the loan-to-value ratio (LTV), and the borrower's creditworthiness (FHFA, 2024). However, as shown by (Bartlett, 2021), it is possible for two different mortgages with the same attributes, origination and maturity dates, but issued by different lenders, to have different interest rates due to the discretionary interest rate component set by the lender. It follows that the interest rate is not simply the result of a deterministic formula of other risk factors, such as LTV and credit score, and therefore may have predictive power on its own. For this reason, the interest rate has been included as a predictor.

As part of the Federal Housing Finance Agency's risk-sharing initiative, Freddie Mac has published loan-level credit data (origination and monthly performance) for all mortgages it has purchased since 1999 to "increase transparency and help investors build more accurate credit performance models" (Freddie Mac, 2024). The data is updated quarterly with the addition of origination data for mortgages purchased between the previous "Origination Cutoff Date" and the current[10], and the update of the monthly performance of existing loans up to the current "Performance Cutoff Date"[11] (Freddie Mac, 2024).

The dataset used for the exploratory data analysis and sample selection includes mortgages originated and purchased by Freddie Mac between the first quarter of 2015 and the last quarter of 2018 and considers their performance through the end of 2023. The dataset contains 2.7 million unique mortgages. A sample of 1.23 million mortgages was extracted from this dataset and used to develop the models. The decision not to use all the available data (from 1999) for modelling was dictated by (1) the exclusion of mortgages that were not considered mature according to the definition of loan maturity below, and (2) computational constraints that forced further reduction of the training set. Nevertheless, the final development sample represents a dataset of real-world mortgage data that is significantly larger than the laboratory datasets commonly used in academic model development and model comparison work, which are in the hundreds or thousands of records. As a result, the absolute and relative performance of the developed models more closely resembles what an industry practitioner might achieve when developing credit scorecards from large volumes of internal and credit bureau data.

All mortgages considered in this paper are part of the Standard Dataset (a subset of Freddie Mac's Single-Family Loan-Level Dataset), which contains origination and performance data on fully amortising fixed-rate single-family mortgages, the type of mortgages eligible for the Single-Family Credit Risk Transfer (CRT) transactions (Freddie Mac, 2024). This category includes

[10] At the time of writing the "Origination Cutoff Date" is 30 September 2023.
[11] At the time of writing the "Performance Cutoff Date" is 31 December 2023.

mortgages with terms of 10, 15, 20, 30 or 40 years with full documentation, i.e. application information verified by the sellers, the finance companies that originated the loans and sold them to Freddie Mac. Given the nature of the data used, the models developed in this paper could be used in practice by anyone holding a portfolio of CRT-like mortgages (investors or lenders) to classify mortgages at risk of default and make informed risk management decisions.

Loan origination information such as borrower credit score, original unpaid principal balance (UPB), loan-to-value ratio (LTV), debt-to-income ratio (DTI), interest rate, or number of units and postcode of the property is available for each of the mortgages. Performance data, on the other hand, includes loan performance information such as current UPB, current estimated LTV (ELTV), number of months remaining to maturity, current interest rate and delinquency status (number of days the borrower is delinquent). These and other variables were used as indicators or predictors of mortgage default in the model development, in their original, aggregated or transformed form. *Appendix A* provides a comprehensive list of all variables used, each with their respective description.

For each mortgage, a snapshot of its performance is published every month until the current Performance Cutoff Date or the occurrence of one of the "termination events": maturity, voluntary redemption, sale, foreclosure charge-off, third party sale, Real Estate Owned (REO) property disposition or re-performing loan securitisation. Termination events are mutually exclusive, i.e. no more than one can occur per mortgage during its lifetime. When a termination event occurs, the performance of the mortgage is no longer updated, and the loan becomes inactive (Freddie Mac, 2024). *Appendix B* presents the definition of each termination event.

As the dataset used includes mortgages originated from 2015 onwards, none of them have reached their maturity date[12]. It follows that all mortgages that were fully repaid were voluntarily prepaid. The prepayment rate of a portfolio is important to a lender because a borrower's voluntary prepayment of a loan reduces the amount of interest earned on that loan and therefore the lender's return. For this reason, some mortgage contracts include prepayment penalties to discourage this behaviour, and models are created by lenders and investors to estimate the prepayment rate of a portfolio of mortgages (Sirignano, Sadhwani, & Giesecke, 2020). However, this paper does not deal with the prediction of the repayment rate, but focuses on the main risk of the loans, the default risk, by developing an interpretable model capable of discriminating mortgages with a higher probability of default from those with a lower probability.

The Target Variable

When developing a probability of default (PD) model, it is necessary to classify each mortgage in the sample as defaulted or non-defaulted. Specifically, in the development of behavioural scorecards, a group of loans is observed at one or more points in time (observation dates) and the

[12] Fully amortising fixed rate mortgages have a minimum term of 10 years. Mortgages with 10-year maturities that were originated in the first quarter of 2015 will mature in the first quarter of 2025. As the current performance cut-off date is 31 December 2023, no mortgages in the dataset have reached maturity.

future performance of each is then analysed over a period of time, typically 6, 12 or 24 months, called the outcome window (Kennedy, Mac Name, Delany, O'Sullivan, & Watson, 2013). Mortgages are classified as performing or non-performing based on whether or not they have defaulted during the outcome window, according to a predetermined definition of default. The classification into bad and good results in a binary outcome variable with a value of:

- 1 if the mortgage defaulted during the outcome window
- 0 otherwise

It is therefore clear that specific definitions of default event and outcome window are needed to define the target variable.

Definition of Default

The definition of default depends on the delinquency status of the borrower. A borrower is classified as delinquent when they miss an instalment payment on their debt. The delinquency status is defined by the number of days past due (DPD), often grouped in 30-day bins, and reverts to current once each outstanding payment is made. A mortgage can be classified as defaulted if, during the outcome window, the borrower's delinquency status reaches a certain threshold (e.g. 60, 90, 180+ DPD) or if one of the following termination events occurs: short sale, charge-off, third-party sale, REO disposition. These termination events occur as a result of foreclosure, which is the legal process by which the lender seeks to recover the outstanding unpaid balance of the loan from a borrower who has stopped making payments by selling the collateral of the loan (the real estate property in the case of mortgages). Foreclosure is therefore an objective indicator of delinquency, but it is too rare an event to be the only one used.

To create a delinquency-based definition of default, it's necessary to establish the threshold of days past due (DPD) beyond which most loans will not recover and therefore result in a loss to the lender. The Basel II Accord defines "default" as 90 DPD and considers this to be the threshold at which the borrower is unlikely to repay the debt in full. 90 DPD is also the most widely used definition of default in credit risk modelling (Siddiqi, 2017). Basel II also uses the "ever bad definition" of default, in which the mortgage is classified as defaulted if it reaches 90 DPD at any time during the outcome window, instead of the "current bad definition" of default, which only considers the delinquency status at the end of the outcome window (Kennedy, Mac Name, Delany, O'Sullivan, & Watson, 2013). This paper uses a definition of default in line with the Basel II Capital Accord:

90+ DPD or short sale/charge-off/third-party sale/REO at any time during the outcome window.

Bad Rate Development Curve

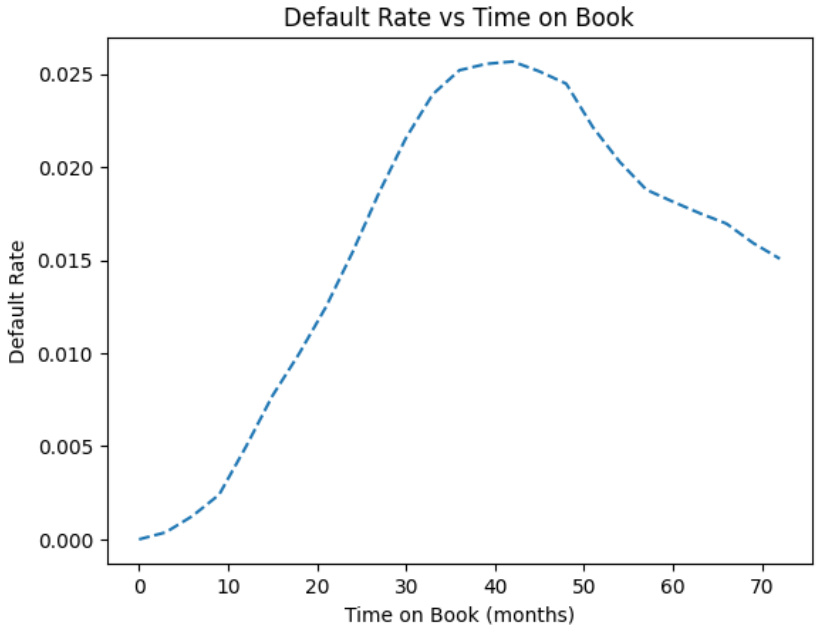

Figure 2: Bad Rate Development Curve

The development sample should be selected from a period when the loan population is considered "mature", i.e. when the loan default rate has stabilised (Siddiqi, 2017). This serves to minimise the likelihood of misclassification due to underestimation of the portfolio default rate. The time to maturity depends on the definition of default and the type of product. Typically, mortgages take four to five years to mature (Siddiqi, 2017). In addition, a looser definition of default, such as 30 DPD, allows loans to mature more quickly than a stricter definition, such as 90 DPD or foreclosure. However, there is a trade-off between maturity and the relevance of the development sample. Older loans are more likely to belong to a mature cohort, but may be outdated due to economic, demographic and policy changes, i.e. they may not reflect the current and future loan population (Siddiqi, 2017). The optimal choice is therefore to select the development sample from the most recent mature loans.

The time it takes for mortgages to mature can be determined analytically by plotting the "default rate development curve" (Siddiqi, 2017). At one or more points in time, the default rate is observed for each vintage, in this case each origination/acquisition quarter, of the mortgages in the portfolio from the beginning of 2015 to the end of 2021. This gives the default rate by time on book (time between origination/acquisition and observation date) for each vintage. For each vintage, the default rate is measured at the end of each of the eight quarters in 2021 and 2022. For each time-on-book value, there are therefore a maximum of eight default rates corresponding to successive vintages. The average default rate is then plotted against the time-on-book to construct the default rate development curve shown in *Figure 2*. Taking the average over several consecutive vintages allows us to smooth out the effect of economic cycles and other factors that may affect the quality

of mortgages originated in a particular vintage (Siddiqi, 2017). The default rate development curve shows that the average default rate begins to decline after about 40 months on the books. This means that mortgages that have been outstanding for at least 40 months can be considered to belong to a mature cohort and can therefore be included in the development sample. In this paper, mortgages originated and purchased by Freddie Mac between the beginning of 2015 and the end of 2018 were sampled, as they were considered mature at the observation dates.

Independent Variables

The selection of predictor variables to be included in the dataset was determined by the availability of such variables in the original data (Freddie Mac, 2024) and their predictive power according to the academic literature. Some of the variables in the original data, such as the maturity date, the name of the seller and servicer, or the Loan Sequence Number of the mortgages, were removed because they were not considered important in discriminating between high and low risk mortgages. New variables were created by transforming and combining existing variables. For the selection of predictive variables, reference was made to the list of "most predictive" and most used variables for the development of retail credit risk scorecards (for mortgage portfolios) provided by (Siddiqi, 2017). The variables considered in this paper can be divided into the following categories:

- Credit Bureau Data: FICO Credit Score.
- Internal/Financial Data: loan-to-value, debt-to-income, worst delinquency in the last 12 months, number of 30/60/90+ DPD in the past 12 months, etc.
- Collateral Data: property type, number of units, first time homebuyer, occupancy status etc.
- Economic Variables: unemployment rate, inflation rate, house price index, median annual income, national 30-years fixed-rate mortgage average rate.

The information contained in the independent variables is available at the time of origination (e.g. credit score, LTV, DTI) or at the time of observation (e.g. estimated LTV, current interest rate, worst delinquency in the last 12 months). Some transformed variables are (1) the difference between the current mortgage rate and Freddie Mac's national mortgage rate, (2) the payment ratio, i.e. the proportion of the original UPB that has been repaid, (3) the balance per month, i.e. the approximation of the principal obligation per month until maturity. These variables are considered to be predictive because (1) an interest rate higher than the national average may encourage strategic defaults, (2) if the mortgage is largely paid off, it is less convenient to default and lose the property used as collateral, (3) a high balance per month indicates significant pressure on the borrower's monthly salary that could lead to default.

Local Economic Variables

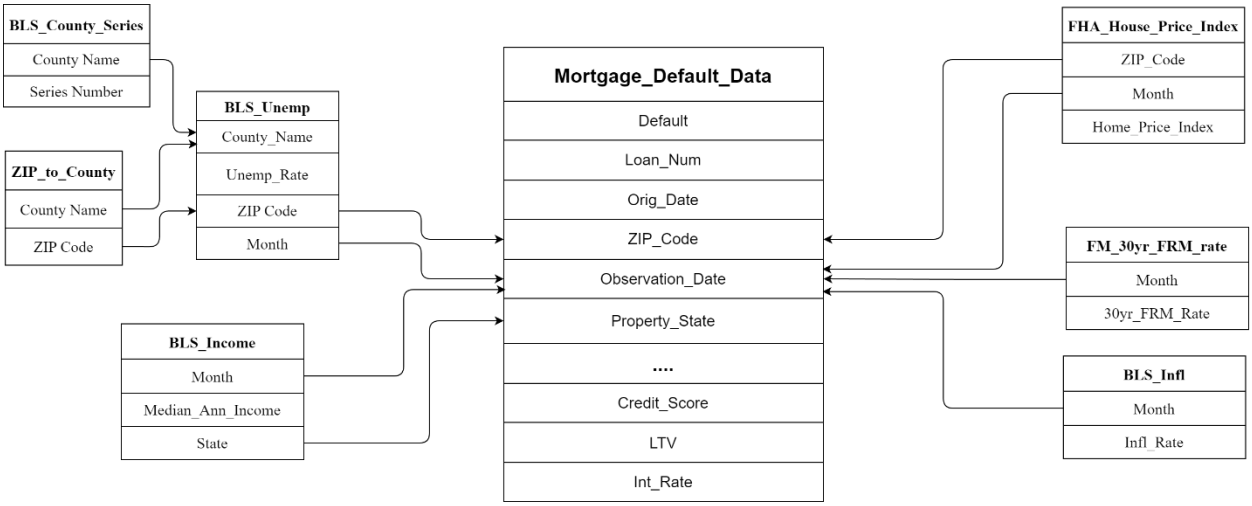

Figure 3: Economic Variables Merging Scheme

Following the assumption that past performance is an indicator of future performance, when developing scorecards it is necessary to select the development sample from a normal business period so that it is representative of the population of loans expected in the portfolio in the future (Siddiqi, 2017). This makes the scorecards more robust and their probabilistic predictions more accurate. It is therefore necessary to consider the effects of seasonality and the economic/business cycle in order to create a representative development sample. In this paper, two techniques have been used to account for the periodic effects of external factors. To counter the effect of seasonality, the development sample was constructed using staked sampling, i.e. the mortgage portfolio was observed at different dates, each with equal length performance and outcome windows (Siddiqi, 2017), as shown in *Figure 4*. To counteract the effect of the business cycle, national and local economic variables down to 3-digit postcode level were included as independent variables. Furthermore, (Sirignano, Sadhwani, & Giesecke, 2020) show the importance of macroeconomic variables in predicting mortgage risk after controlling for the effect of other factors considered to be significant predictors (e.g. LTV, credit score, DTI). In particular, they develop a deep learning model for predicting the state transition matrix (probability of transition between different payment/delinquency states: current, 30 DPD, 60 DPD, 90 DPD, foreclosure, REO, paid off) for a large portfolio of mortgages (120 million). They then calculate the economic significance of each independent variable by the "average magnitude over the data of the derivative of a fitted transition probability with respect to the variable" (Sirignano, Sadhwani, & Giesecke, 2020). The state unemployment rate turns out to be the variable with the greatest explanatory power. Although this result is the result of an Explainable AI (uses derivative) method applied to a black box model (neural network), it suggests that the economic conditions faced by a borrower affect his probability of default. Based on this consideration, the following economic variables were

included in this paper: (1) unemployment rate at the county level[13], (2) median annual income at the state level[14], (3) house price index at the three-digit zip code level[15], (4) annual inflation rate at the national level[16], (5) average 30-year fixed-rate mortgage rate at the national level[17]. Data for each of these economic variables are published at different frequencies and with different time lags from the reference period: (1) unemployment rate: monthly frequency, released at the end of the following month or at the beginning of the second month[18], (2) median annual income: annual frequency, released in May or April of the following year[19], (3) house price index: quarterly frequency, published two months later[20], (4) inflation rate: monthly frequency, published by the mid of the following month[21], (5) U.S. 30-year fixed-rate mortgage average rate: weekly frequency, published on Thursdays[22]. Since the economic data are published by different institutions (Bureau of Labour Statistics, Federal Housing Finance Agency, Freddie Mac), with different frequencies and release schedules, it was necessary to standardise the data to merge the economic variables with the other independent variables, avoiding look-ahead bias, i.e. using only the data available at each observation date. For example, considering a sample of mortgages observed at the end of March 2021 (end of the first quarter), the default prediction uses data on the unemployment rate for January 2021, the median annual income for May 2020, the house price index for January 2021, the inflation rate for February 2021 and the average 30-year fixed-rate mortgage rate for the third week of March 2021. The diagram in *Figure 3* shows the process of merging the economic data with the mortgage data, based on the frequency, release schedule and geographical scale of the economic variables. Data at a finer scale (e.g. postcode) is preferable as it more accurately captures the economic scenario faced by the borrower. For this reason, each economic variable was considered at the finer geographical scale freely available online.

Missing Values

It is common for financial data to contain missing values (Siddiqi, 2017). Some models, such as those based on decision trees, are immune to the presence of missing values in the development set, while others, such as logistic regression or generalised additive models, require complete data sets. Freddie Mac's single-family loan-level dataset also contains missing values for some fields, as shown in *Appendix A*. In cases where the reason for the missing data is not known, one can address it in the development sample by: (1) eliminating all records with missing values for one or more characteristics, this approach is called “complete case analysis” and may result in a very small dataset, (2) eliminating variables or records with an excessive proportion of missing values,

[13] Local Area Unemployment Statistics (LAUS) from the U.S. Bureau of Labor Statistics https://www.bls.gov/lau
[14] Occupational Employment and Wage Statistics from the U.S. Bureau of Labor Statistics https://www.bls.gov/oes
[15] House Price Index from FHFA https://www.fhfa.gov/data/hpi
[16] Consumer Price Index from the U.S. Bureau of Labor Statistics https://www.bls.gov/cpi
[17] Mortgage Rates from Freddie Mac https://www.freddiemac.com/pmms
[18] https://www.bls.gov/lau/laufaq.htm#Q07
[19] https://www.bls.gov/oes/oes_ques.htm
[20] https://www.fhfa.gov/data/hpi#ReleaseDates
[21] https://www.bls.gov/schedule/news_release/cpi.htm
[22] https://www.freddiemac.com/pmms/about-pmms

(3) imputing the missing values using statistical techniques that take into account the value of other records or characteristics (Siddiqi, 2017; Florez-Lopez, 2010). However, these methods assume that the missing values have no value, are random (Florez-Lopez, 2010), and therefore do not contribute to the development of the model. In reality, missing values are often not random (Siddiqi, 2017). For example, an applicant may decide to omit income because it is too low, or credit score may be marked as missing because it is outside the permitted range. The approach recommended in (Siddiqi, 2017) and used in (Sirignano, Sadhwani, & Giesecke, 2020) is to encode the missing values as an additional indicator variable, i.e. a binary variable that takes the value 1 if the value of the variable to which it refers is missing and 0 otherwise. This method recognises the value of missing data and makes it possible to develop a model with accurate predictions without the need for recalibration.

For some variables in the Freddie Mac Single Family Loan-Level Dataset, the reason for missing values is specified in the documentation (Freddie Mac, 2024). The FICO credit score is reported as missing if its value is outside the allowed range of 300-850, the debt-to-income ratio if it is greater than 65%, and the mortgage insurance percentage if it is outside the range of 1%-55%. In these cases, specific indicator variables have been created (Credit Score out of 300-850, DTI > 65%, Mortgage Insurance out of 1%-55%). In cases where the reason was not known, an indicator variable was created for each independent variable with missing values and included in the development dataset.

Development Sample

Performance and Outcome Window

The development of credit risk scorecards involves analysing the past performance of accounts/loans to predict future performance, on the assumption that historical data reflects future behaviour (Siddiqi, 2017). In the case of behavioural scorecards, a sample of loans is observed at one or more points in time, called observation dates. In this paper, a sample of mortgages extracted from Freddie Mac's Single-Family Loan-Level Dataset was observed at eight observation dates, corresponding to the end of each quarter in 2021 and 2022.

The repayment behaviour of the loans is then observed for a period of time following the observation date, known as the outcome window, in order to determine default based on the established definition and to assign the corresponding value of the target variable to each loan in the sample at each observation date. In this work, a 12-month outcome window has been considered, in line with the Basel II regulatory framework, although more recent frameworks such as Basel III and IFRS 9 emphasise the importance of using a window length appropriate to the characteristics of the portfolio being analysed and capable of capturing possible macroeconomic factors that could influence the default rate (Siddiqi, 2017).

The objective of a behavioural scorecard is to correctly predict the target variable (default/non-default) for each loan using the information available at the observation date. This information includes data on the origination and historical performance of the loan, as well as the economic

scenario. Performance data were collected by observing each loan for a period of 12 months prior to the observation date (performance window). The number of times the borrower was 30, 60, or 90+ days delinquent, the maximum delinquency, and the number of modifications in the 12 months prior to each observation date were recorded. The choice of a 12-month performance window was influenced by (Kennedy, Mac Name, Delany, O'Sullivan, & Watson, 2013), which evaluates the performance of a behavioural scorecard (ridge logistic regression) with different lengths of performance and outcome windows on a 7-year dataset of Irish mortgage data. The study shows that using a performance window of 12 months, the classifier achieves the highest average class accuracy for outcome windows of 3, 6 and 12 months.

Training and Test Sets

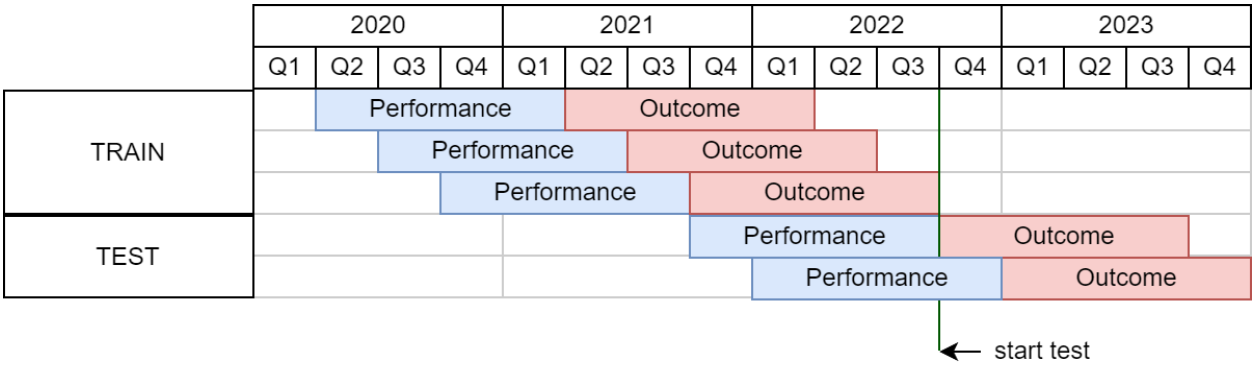

Figure 4: Stacked Sampling: Train-Test Split

The development sample was constructed by observing a portfolio of mortgages originated between the first quarter of 2015 and the last quarter of 2018 on eight observation dates corresponding to the end of each quarter in 2021 and 2022, following the principle of stacked sampling. *Figure 4* shows how the development sample was subsequently divided into a training set and a test set. The training set was used to develop the machine learning models, while the test set was used to estimate and compare their out-of-time performance. It is important to note that the first observation date of the test set coincides with the end of the outcome window of the last observation date of the training set. This approach approximates the realistic situation where a model is developed at the end of the third quarter of 2022, using only the data available at that time, and implemented from then on.

Due to computational constraints, the training set does not contain all mortgages in the portfolio, but a random sample of 500,000 loans drawn at each observation date, for a total of 1.5 million. The test set, on the other hand, contains ~3.55 million mortgages. Using a larger test set than the training set is not a conventional choice but allows for conservative estimates of model out-of-time performance.

Class Imbalance

The resulting training set has an extreme class imbalance, as is common in loan default data. Only 0.96% of the mortgages are classified as defaulted, a total of 14299 loans. However, the number of "bad" loans is greater than 5000, making the dataset suitable for the development of a statistically significant credit scorecard, according to (Siddiqi, 2017). Class imbalance makes the minority class (in this case, defaulted mortgages) more difficult to predict because there are few examples of it, while the classification model could be biased towards the majority class because there are many examples of it in the data set. There are several methods to resolve class imbalance, such as oversampling the minority class, undersampling the majority class, generating synthetic data using techniques such as SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In this paper, however, following (Rahmani, Parola, & Cimino, 2024), classes are not rebalanced. Since different models are compared on the same dataset, if class imbalance affects all classifiers equally, the relative performance remains unchanged, whereas if some classifiers are more robust to class imbalance, this property should be considered in the comparison.

However, it is important to use performance measures that are robust to class imbalance when evaluating models. For example, accuracy is not an appropriate metric in the case of imbalanced classes. A naive classifier that labels all mortgages as "non-default" would achieve an almost perfect accuracy of 99.04% on the train set, while misclassifying all defaulted loans, which are the ones we are most interested in correctly classifying. For this reason, the models were compared using metrics that are robust to class imbalance, such as AUC, Brier score and average precision.

Model Development

Sparse Generalized Additive Model

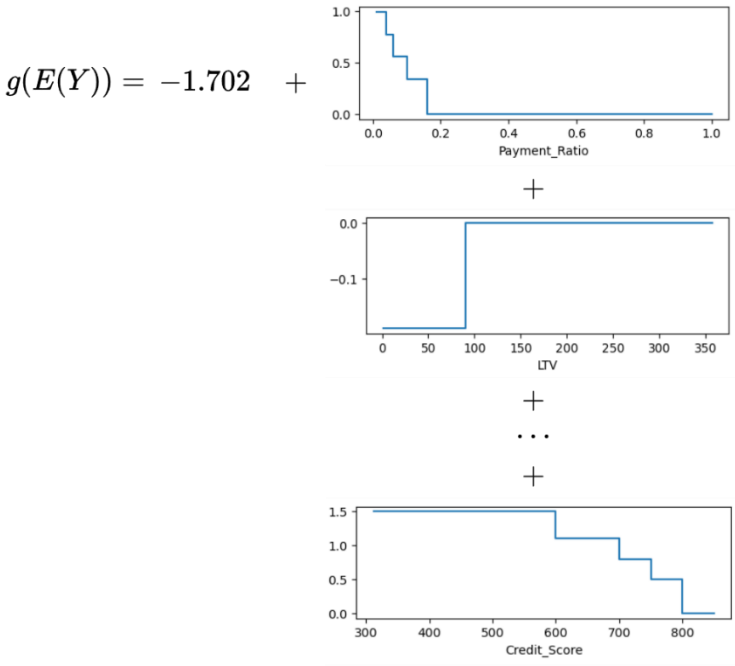

Figure 5: Example of Generalized Additive Model with Step Component Functions

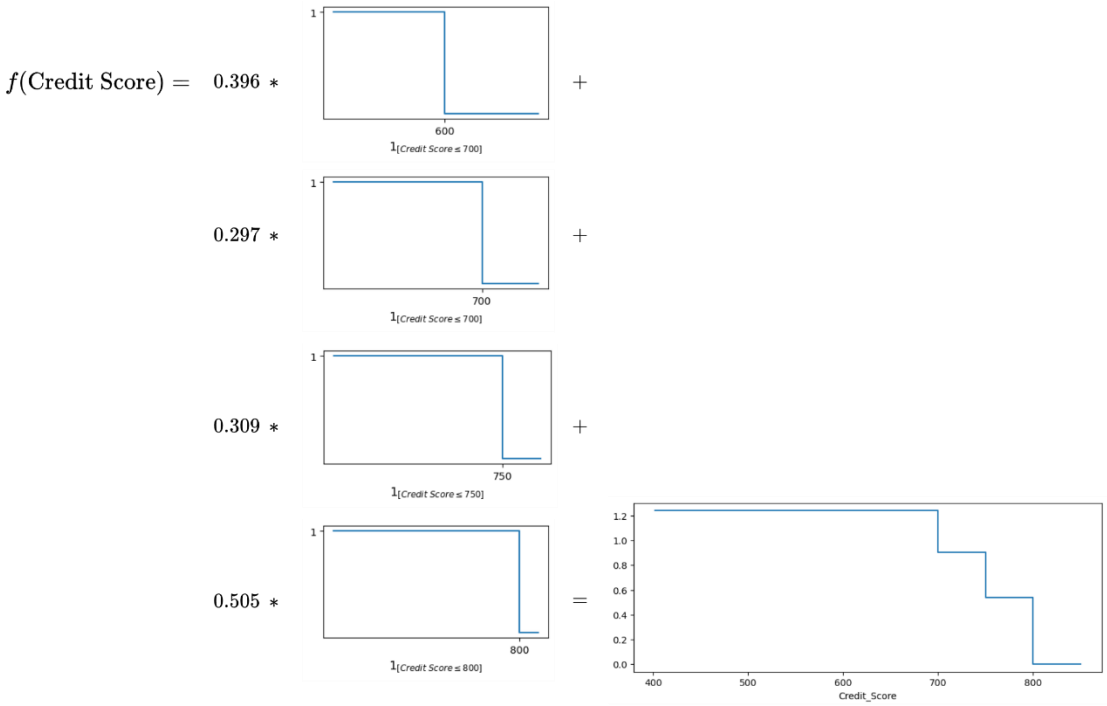

Figure 6: Example of Component Function

A Sparse Generalised Additive Model (GAM) was chosen as the intrinsically interpretable model, the algorithm for which is presented in (Liu, Zhong, Seltzer, & Rudin, Fast Sparse Classification for Generalized Linear and Additive Models, 2022) .

A GAM is a model in which the expected value $E(Y)$ of the response variable Y depends linearly, via a link function g, on functions $f_1, \dots, f_p$, called component functions, of the independent variables $x_1, \dots, x_p$ (Tibshirani & Hastie, 1986).

$$g\big(E(Y)\big) = \alpha_0 + \sum_{j=1}^{p} f_j(x_j) \tag{1}$$

When the generalised additive model is used for classification with a binary target variable Y, the logit function can be used as the link function. In this way, the real score $g\big(E(Y)\big)$ obtained from the sum of the intercept α_0 and the component functions can be converted into probability estimates with values in $[0,1]$.

$$g\big(E(Y)\big) = log\ \frac{P\big(Y=1\big|x_1,\dots,x_p\big)}{P\big(Y=0\big|x_1,\dots,x_p\big)} \ \rightarrow\ p = P\big(Y=1\big|x_1,\dots,x_p\big) = \frac{e^{g\left(E(Y)\right)}}{1+\ e^{g\left(E(Y)\right)}} \tag{2}$$

The functions of the independent variables are unknown and must be estimated. They may be parametric, non-parametric or semi-parametric. The freedom in the choice of the functional form allows the modelling of non-linear relationships between the individual characteristics and the dependent variable. In this paper, following (Liu, Zhong, Seltzer, & Rudin, 2022), piecewise (step) functions were used, obtained by combining decision stumps by boosting, as shown in *Figure 6*. Decision stumps are one-level decision trees in which the root node is directly connected to the leaves. They make decisions based on a single feature. For binary classification the tree has two leaves and for continuous variables it can be represented by a single step function. They are therefore weak learners and are often used as components of bagging or boosting ensembles such as AdaBoost (Freund & Schapire, 1997). In the boosted stumps algorithm, step functions are created by choosing a feature (independent variable) and a threshold in the case of numerical variables, or a value in the case of categorical variables, at each iteration, creating a decision stump and then assigning a coefficient to it. Finally, all the stumps are sorted by feature and added to create a step function for each feature, as in *Figure 6*. In this work, the thresholds for each numerical variable were chosen a priori. For each variable with more than 20 unique values in the development sample, the values corresponding to the 5th to 100th percentiles were considered as thresholds and a binary variable of type $1_{[NumVar \leq threshold]}$. For some variables, the thresholds were rounded to increase their interpretability. For example, for credit score, the thresholds were rounded to the nearest 50th, as a variable of type $1_{[Credit\ Score \leq 750]}$ is more interpretable than $1_{[Credit\ Score \leq 762]}$. . In the case of numeric variables with a maximum of 20 unique values, all values were used as thresholds, while for categorical variables, a binary variable of type

$1_{[CatVar = category]}$ was created for each category. The resulting development set contains 512 binary variables. Encoding the features before training the model allows us to (1) make the model more interpretable by choosing thresholds that are common in credit risk models, (2) reduce the risk of overfitting by choosing thresholds that are too granular, and (3) impose monotonicity constraints on certain features. The latter is particularly useful in credit risk modelling where certain features are known to have a positive or negative impact on credit risk. For example, default risk should be monotonically decreasing in credit score and monotonically increasing in LTV, as shown in *Figure 5*.

The resulting generalised additive model can then be expressed as a summation of the individual stumps as follows:

$$g\big(E(Y)\big) = \sum_{j=1}^{p} \sum_{threshold\ l} \sum_{t=1}^{T} \alpha_t\, 1_{[h_t = 1_{[x_j \le \theta_l]}]}\, 1_{[x_j \le \theta_l]} = \sum_{t=1}^{T} \alpha_t\, h_t(x)$$

where θ_l is the l^{th} threshold of the feature x_j s selected at iteration t of the boosting algorithm. $1_{[h_t = 1_{[x_j \le \theta_l]}]}$ an indicator variable that takes the value 1 if the stump $1_{[x_j \le \theta_l]}$ is selected at iteration t and 0 otherwise.

To fit a generalised additive model by classification, it is necessary to solve the following optimisation problem, whose objective function is called the "loss function":

$$min_w \sum_{i=1}^{n} l(\boldsymbol{\alpha}, \boldsymbol{x}_i, y_i) + Regularization)$$

Following (Liu, Zhong, Seltzer, & Rudin, 2022), the logistic loss was used, which returns well-calibrated probability estimates via (2):

$$l(\boldsymbol{\alpha}, \boldsymbol{x}_i, y_i) = log\left(1 + e^{-y_i(\boldsymbol{\alpha}^T \boldsymbol{x}_i)}\right)$$

where $x_i \in R^p$ is the i^{th} observation and $y_i \in \{-1; 1\}$ is the label of the i^{th} observation. In order for the minimisation problem to produce a generalised additive model, the features $x_1, \dots, x_p$ were transformed into binary variables (stumps) of the type $1_{[NumVar \le threshold]}$ or $1_{[CatVar = category]}$, as explained above. The vector $\boldsymbol{\alpha}$ contains the coefficients of the stumps, which are zero if the stump is not included in the model.

Following (Liu, Zhong, Seltzer, & Rudin, 2022), a l_0 penalty, which represents the number of non-zero coefficients $\|\boldsymbol{\alpha}\|_0$, has been included as a regularization term. A model with a high number of total stumps (whose coefficients are non-zero) will therefore be penalised more. The λ_0 parameter, called the regularisation constant, controls the relative importance of the l_0 penalty to the logistic

loss in the loss function. A l_2 penalty term $\|\boldsymbol{\alpha}\|_2^2$ with a small constant λ_2 is added to speed up convergence during optimisation. The resulting problem looks as follows:

$$min_{\alpha} \sum_{i=1}^{n} log\left(1 + e^{-y_i(\boldsymbol{\alpha}^T \boldsymbol{x}_i)}\right) + \lambda_0\|\boldsymbol{\alpha}\|_0 + \lambda_2\|\boldsymbol{\alpha}\|_2^2 \quad (3)$$

Including the l_0 makes it possible to create sparse generalised additive models with a small number of stumps. This makes such models interpretable, as the component functions can be visualised in a single plot, as shown in *Figure 8*. At the same time, however, the penalty l_0 penalty makes problem (3) NP-Hard. (Liu, Zhong, Seltzer, & Rudin, 2022) introduce a best subset search algorithm with quadratic cuts and dynamic feature ordering to solve (3) and generate sparse and accurate generalised linear or additive models from large datasets 2 to 5 times faster than previous techniques, even with many highly correlated features.

The model was implemented using the fastSparseGAMs[23] (Liu, Zhong, Seltzer, & Rudin, 2022; Hussein & Rahul, 2020; Dedieu, Hazimeh, & Mazumder, 2021) Python package, which was developed using the L0Learn package (Hazimeh, Mazumder, & Nonet, 2022) and is based on the algorithms introduced by (Liu, Zhong, Seltzer, & Rudin, 2022), for solving L0-regularized learning problems in sparse additive models. The model was fitted using logistic loss, the CDPSI algorithm and iterating over a set of up to 150 decreasing λ_0 values until a maximum of 40 attributes (binary variables as described above) were included in the model. As λ_0 decreases, so does the strength of the l_0 penalty and thus the number of variables in the model, called "support size". The fastSparseGAMs documentation recommends using a small fraction of $\min(n, p)$ as maximum support size. The fitted model has 31 attributes, which makes it easy to visualise and interpret, as can be seen in *Figure 8*. The full model is shown in *Appendix C*.

Tree-Based Models

Two tree-based ensemble models have been developed as black-box benchmarks for the interpretable generalized additive model: Random Forests and Gradient Boosted Decision Trees. These models are black box because they are constructed from the combination of tens, hundreds or thousands of decision trees, making them uninterpretable. At the same time, however, they have proven to be the state of the art for classification tasks involving tabular data. Their performance surpasses that of any other model architecture, including neural networks, when developed on structured data, as shown in the massive benchmarking exercise presented in (Grinsztajn, Oyallon, & Varoquaux, 2022). Since most credit risk modelling problems, including mortgage default prediction, involve tabular data, it is reasonable to consider the performance of tree-based ensemble models as the best achievable and therefore the one to aim for.

A Random Forest (Ho, 1995) is an ensemble of decision trees whose predictions are averaged. In classification tasks, each input is assigned the class chosen by the majority of trees. This model

[23] Code available at https://github.com/ubc-systopia/L0Learn/tree/master/python

architecture was developed to overcome the problem of overfitting in decision trees. By training multiple small decision trees, which are weak learners with high bias and low variance, and averaging their predictions, it is possible to reduce the variance of the overall model without significantly increasing its bias. Other techniques are also used to avoid overfitting, such as training trees on random subsamples of the training set rather than the entire data set and using randomly selected subsets of features. In the benchmarking of credit scoring classification algorithms presented by (Lessmann, Baesens, Seow, & Lyn C., 2015) Random Forest achieves the best performance among homogeneous ensembles, so the authors recommend using it as a benchmark against which to compare new classification algorithms for credit risk scoring.

A Gradient-Boosted Decision Tree is an ensemble of simple decision trees, i.e. weak learners, that are iteratively combined to create a strong learner. Starting from a naive model, each iteration of the gradient boosting algorithm is trained to correct the errors of its predecessor. To do this, each weak learner is trained using so-called pseudo-residuals, i.e. the negative gradient of the loss function with respect to the current model, evaluated at the predictions of the previous model, instead of the initial target variable. Gradient boosting makes it possible to minimise any differentiable loss function and produce accurate and robust models. Gradient-boosted decision trees, such as AdaBoost (Freund & Schapire, 1997) or XGBoost (Chen & Guestrin, 2016), were not included in (Lessmann, Baesens, Seow, & Lyn C., 2015) benchmarking, but are expected to perform better or equal to Random Forests.

In this work, a random forest classifier implemented using the scikit-learn Python package[24] and a regularised gradient boosted decision tree model using XGBoost (eXtreme Gradient Boosting) (Chen & Guestrin, 2016), implemented using the xgboost open-source Python package[25] were developed. Both models were trained and tested on the same training and test set as the generalised additive model. The hyperparameters of the models were optimised by cross-validated grid-search over a parameter grid. Due to computational constraints, the range of hyperparameters tested is not particularly wide, but was constructed with the aim of reducing overfitting and facilitating training in the presence of highly imbalanced data. Monotonicity constraints were imposed on key variables for both models: the estimated probability of default is monotonically increasing in LTV, combined LTV, debt-to-income, interest rate and current delinquency status, while it is monotonically decreasing in credit score and payment ratio. The choice of key variables is based on both business logic and interpretation of the sparse generalised additive model. Finally, model calibration was performed. Tree-based models by their nature do not provide probability estimates, although techniques exist to derive predicted probabilities from them. However, when looking at the calibration plot, it was noted that the predicted probabilities of the random forest did not accurately reflect the distribution of the classes. This is not the case for XGBoost, which provides well-calibrated probability estimates. The random forest classifier was then calibrated using the

[24] Scikit-learn RandomForestClassifier https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[25] XGBoost documentation available at https://xgboost.readthedocs.io/en/stable/index.html

CalibratedClassifierCV[26] method from scikit-learn. This method divides the training set into 5 stratified folds, each with approximately the same percentage of samples for each class. For each of the 5 folds, the classifier is trained on the remaining 4 and the probabilities are calibrated on the fifth. The probabilities are calibrated by fitting a regressor, called a "calibrator", which maps the predictions of the classifier to probabilities that reflect the class distribution in the dataset. Isotonic regression was used as the calibration method. The use of stratified k-fold cross-validation means that the classifiers and calibrators are trained on different subsamples to reduce bias. For prediction, the calibrated probabilities are averaged over the 5 individually calibrated classifiers. Calibration can also improve model performance, particularly in terms of probability-based metrics such as the Brier score.

[26] https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html

Empirical Results

Performance Evaluation

There are three types of performance measures (Baesens, et al., 2003). There are those that assess the discriminatory ability of the model, those that measure the accuracy of probability predictions and those that express the correctness of categorical predictions. These groups of metrics express different aspects of model performance and at least one metric of each type should be included in the model comparison. Three different performance metrics were used in this paper: AUC, Scaled Brier Score and Average Precision.

The AUC is the area under the receiver-operating characteristic (ROC) curve. The ROC curve is obtained by plotting the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis for each possible probability threshold, i.e. the value used for classification, such that if the predicted probability is greater than the threshold, the input is labelled 1 and 0 otherwise. The AUC is the area under the ROC curve and allows comparison of models. It can be interpreted as the probability that the model will rank a randomly chosen positive case higher than a randomly chosen negative case. It is a measure of discriminatory ability and takes values in the range $[0,1]$. A perfect classifier has $TPR = 1$ and $FPR = 0$, so its ROC curve will have a square shape and $AUC = 1$. A random classifier, on the other hand, has $AUC = 0.5$. The goal is to develop a model with an AUC that is as close as possible to 1 and always greater than 0.5.

The Brier Score (BS) is the mean-squared error of the predicted probabilities with respect to the actual labels. It assesses the accuracy and calibration of the estimated probabilities. To improve interpretability, the Scaled Brier Score was calculated, where the BS of the model is normalised to the BS of a non-informative model that always predicts the observed default probability. In this way, the Scaled BS has a range $[0,1]$, where 1 indicates a perfect model and 0 indicates a non-informative model.

$$BS = \sum_{i=1}^{N} (y_i - \hat{p}_i)^2 \qquad BS_{scaled} = 1 - \frac{BS}{BS_{non-inf}}$$

In the presence of class imbalance, the AUC can be misleading, as the model may simply rank the abundant and easy to classify negative cases below the positive cases, resulting in a high AUC, while struggling to correctly classify positive cases. In these cases, it is preferable to use the precision-recall curve instead of the ROC curve. By plotting precision against recall for each possible threshold, it is possible to assess how well the model identifies the rare positive cases. To summarise the precision-recall curve and compare models, the area under the curve was approximated using the average precision (AP), the average precision P a at each threshold n, weighted by the increase in recall R from the previous threshold $n - 1$. The average precision also takes values in the range $[0,1]$, with 1 representing maximum precision.

$$AP = \sum_{n=1}^{N} (R_n - R_{n-1}) P_n$$

The following table shows the values of the performance metrics on the test set for each of the models.

	AUC	Scaled Brier Score	Average Precision
Sparse GAM	0.922	0.211	0.342
Random Forest	0.917	0.133	0.267
XGBoost	0.925	0.226	0.357

Despite being sparse, the generalised additive model performed in line with the black box models. The three models show very similar performance, with XGBoost achieving slightly better results. It is therefore clear that the increase in complexity has not led to a corresponding increase in performance. This is evidence that there is no trade-off between interpretability and accuracy in this classification problem.

Model Interpretation

The structure of the sparse generalised additive model allows for easy interpretation. *Figure 8* shows the full model. For categorical variables, all attributes are included, while for continuous variables, the component functions are plotted. The intercept and coefficients have been rounded to three decimal digits. *Appendix C* reports the full model with all 31 attributes and the original coefficients.

The model respects business logic, although no monotonicity constraints have been imposed. The score represents the risk of default and can be mapped to probabilities. The total score is monotonically increasing in loan-to-value, debt-to-income, current delinquency status, number of 60-day delinquencies in the past year and maximum delinquency status in the past year. The risk of default is known to increase as loan-to-value and debt-to-income increase. In addition, borrowers who are already 30 DPD or 60 DPD are more likely to default. Repayment behaviour during the performance window and, in particular, the frequency and severity of delinquencies also influence the risk of future default. The score decreases monotonically with credit score, number of borrowers, state median annual income, original unpaid principal balance and repayment ratio. The FICO credit score has proved to be a good indicator of creditworthiness. Mortgages with more than two borrowers, higher principal balances and in states with high median incomes are less risky. The presence of multiple borrowers can help distribute risk. Borrowers with higher incomes are more likely to be able to meet the monthly repayments. Larger loan amounts are required for higher value properties purchased by more affluent borrowers. The importance of the repayment rate is interesting. It is evident how the fact that more than ~20% of the UPB has already been repaid makes default much less likely. Looking at the categorical variables, characteristics related

to interventions to address borrowers' financial hardship are associated with an increase in the risk of default. Repayment assistance, mortgage modification and payment deferral are all intended to provide a borrower with a high risk of default with some payment relief or with an opportunity to cure a delinquency. The fact that they have a positive coefficient may indicate that they are not very effective. The trial period, on the other hand, turns out to be a very effective risk mitigation plan. In addition, borrowers who obtain a mortgage to buy their first home present a higher risk.

Among the numerical variables, two deserve closer analysis: the age of the loan and the average interest rate on the US 30-year fixed-rate mortgage. The model assigns higher risk to mortgages originated between 29 and 13 months before the observation date. In addition, the risk of default appears to increase as the current average mortgage rate decreases. The mortgages considered in the development set are originated between the beginning of 2015 and the end of 2018 (grey area in *Figure 7*), so their loan age should be at least 27 months at the observation dates (dashed lines in *Figure 7*). The loan age is reset in the event of a loan modification. It is therefore clear that this variable acts as a proxy for the "loan modification" variable. One solution might be to remove the age of the loan from the independent variables. Looking instead at the historical data for the average US 30-year fixed-rate mortgage in *Figure 7*, we can see that from the beginning of 2019, the interest rate began a descent from historical highs until it reached lows at the observation dates and began a sharp rise in 2022 (blue area in *Figure 7*). Borrowers who were granted mortgages with an average interest rate of ~4% may have chosen to strategically default at the end of 2021, taking advantage of lower interest rates and driven by the prospect of impending increases.

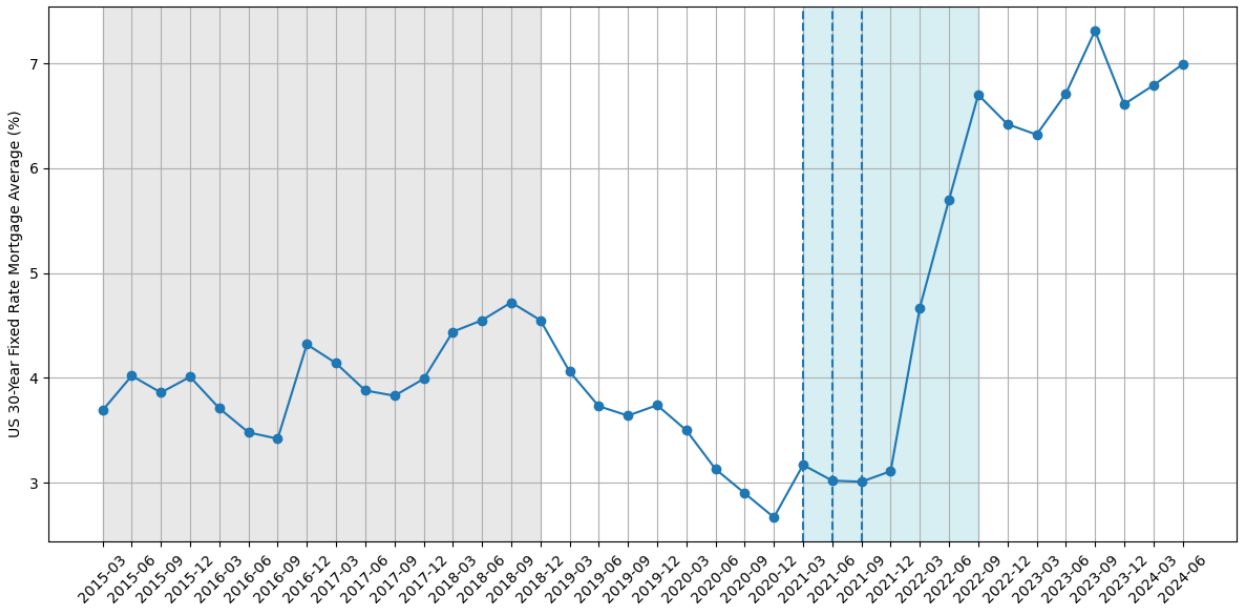

Figure 7: U.S. 30-year Fixed-Rate Mortgage Average 2015-2024

$$
\begin{aligned}
\text{score} = & -1.703 \\
& -0.251 \times 1_{\text{Occupacy\_Status=Second Home}} \\
& -0.382 \times 1_{\text{Channel=Retail}} \\
& +0.110 \times 1_{\text{Loan\_Purpose=No Cash-Out Refinance}} \\
& +1.087 \times 1_{\text{Assistance\_Status=Repayment}} \\
& -1.203 \times 1_{\text{Assistance\_Status=Trial Period}} \\
& +0.469 \times 1_{\text{Modification\_Flag=Prev. Period}} \\
& +0.469 \times 1_{\text{Payment\_Deferral\_Flag=Yes}} \\
& +0.223 \times 1_{\text{Payment\_Deferral\_Flag=Prev. Period}} \\
& +0.218 \times 1_{\text{First\_Time\_Home\_Buyer\_Flag=Yes}} \\
& +0.193 \times 1_{\text{Prepay\_Penalty=Yes}} \\
& +0.174 \times 1_{\text{Num\_Units=Missing}}
\end{aligned}
$$

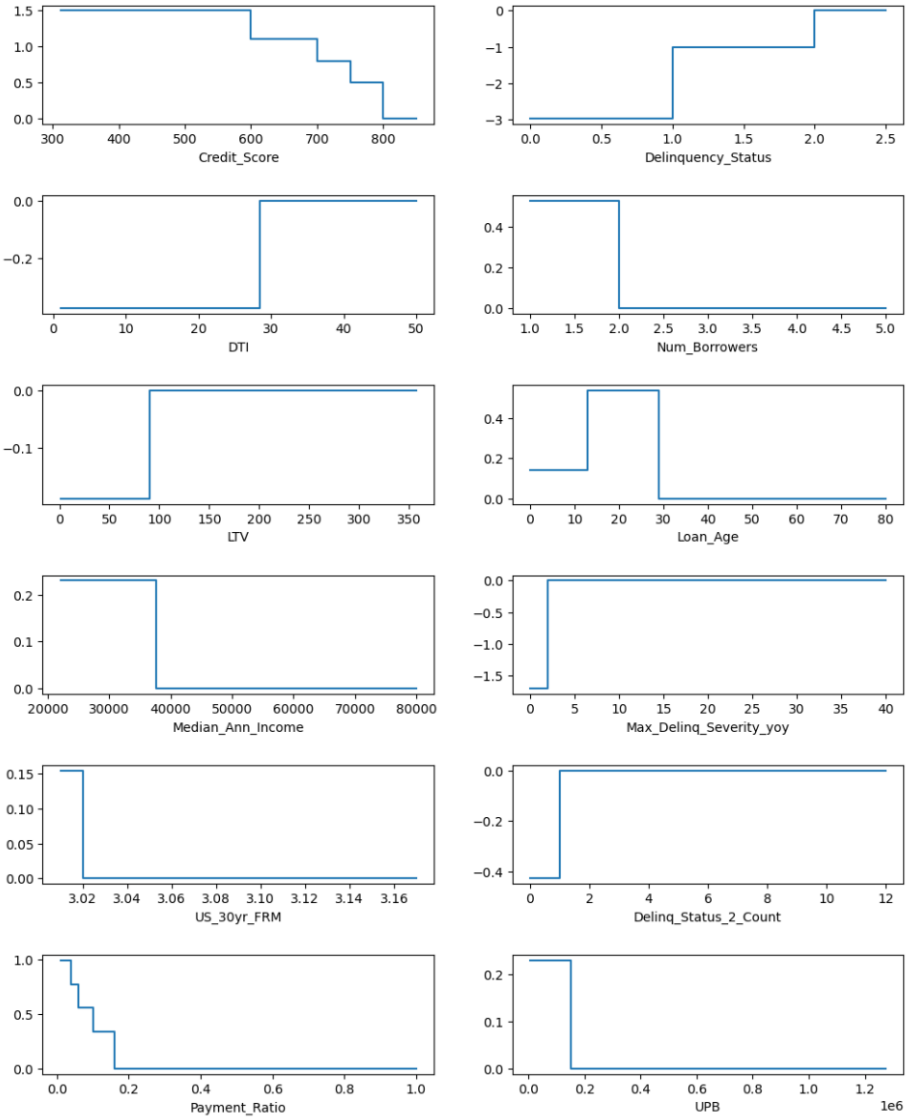

Figure 8: Sparse Generalized Additive Model

Conclusion

This paper shows that increasing the complexity of machine learning models often does not improve their performance, especially for tabular data. An interpretable generalized additive model, developed using a state-of-the-art optimisation algorithm (Liu, Zhong, Seltzer, & Rudin, 2022), achieved the same performance as tree-based ensembles on a large dataset of mortgage default data, while maintaining desirable qualities such as sparsity, monotonicity and decomposability. This result should motivate practitioners and academics to focus more on interpretable models for credit risk modelling, without falling into the interpretability-accuracy trade-off myth. Future work should develop new algorithms, new model structures and new techniques to produce increasingly accurate and interpretable models. Larger benchmarking studies should be conducted using real-world datasets to confirm that sacrificing interpretability does not improve performance, but often degrades it. Finally, domain experts should be more involved in academic studies to allow a clear definition of interpretability for each specific problem.

Appendix A: Variables Description

Variable Name	Original Name	Description	Values
Orig_Date	ORIGINATION DATE	Mortgage Origination Date. Inferred: 2 months on average between Orig_Dated and First_Payment_Date	YYYY-MM
Credit_Score	CREDIT SCORE	Borrower's creditworthiness score provided by FICO.	300-850 If outside of the range then missing
First_Payment_Date	FIRST PAYMENT DATE	Date of first scheduled mortgage payment.	YYYY-MM
First_Time_Homebuyer_Flag	FIRST TIME HOMEBUYER FLAG	The borrower(s) is/are buying the property, will live in it and has/have not owned another residence in the last three years.	Y = Yes N = No Nan = Not Available or Not Applicable
Maturity_Date	MATURITY DATE	Month of the final monthly payment as in the original mortgage note.	YYYY-MM
Metropolitan_Area	METROPOLITAN STATISTICAL AREA (MSA) OR METROPOLITAN DIVISION	MSA has at least one urbanised area with 50k+ inhabitants. MSA with a 2.5+ population may be divided into counties (Metropolitan Divisions).	5-digit MSA/Division code NaN = Not an MSA/Division or Unknown
MI_Percentage	MORTGAGE INSURANCE PERCENTAGE	Original percentage of the loan's default loss covered by the insurer.	1% - 55% 0% = No MI NaN = outside range
Num_Units	NUMBER OF UNITS	One-, two-, three- or four-unit property.	1-4 NaN = Not Available

Occupancy_Status	OCCUPANCY STATUS	Occupancy status of the property.	P = Primary Residence I = Investment Property S = Second Home NaN = Not Available
CLTV	ORIGINAL COMBINED LOAN TO VALUE	Original loan amount plus any secondary mortgage loan amount (all secured loans on the property) divided by the lesser of the appraised value or purchase price of the mortgaged property.	2018Q1 and prior: 6% - 200% 2018Q1 and later: 1% - 998% NaN = CLTV < LTV or Not Available
DTI	ORIGINAL DEBT-TO-INCOME RATIO	Sum of the borrower's monthly debt payments (including the mortgage payment at purchase) divided by the total monthly income used to underwrite the loan.	0% - 65% NaN = DTI > 65%, Not Available
UPB	ORIGINAL UPB	Unpaid principal balance: the portion of the loan that has not been paid at the reporting date.	
LTV	ORIGINAL LTV	Original mortgage loan amount at note date divided by the lesser of the appraised value of the mortgaged property at note date or its purchase price.	2018Q1 and prior: 6% - 105% 2018Q1 and later: 1% - 998% NaN = out of range
Int_Rate	ORIGINAL INTEREST RATE	Interest rate of the mortgage at origination.	
Channel	CHANNEL	Broker: person/entity who specialises in lending and receives a commission for matching borrowers and lenders. Correspondent: entity that sells the mortgage it originates to other lenders under a specific	R = Retail B = Broker C = Correspondent T = TPO (Third Party Originator) Not Specified

		agreement. commitment or as part of ongoing relationship Retail: loan originated, underwritten and funded by a lender.	NaN = Not Available
Prepay_Penalty	PREPAYMENT PENALTY MORTGAGE FLAG	PPM is a mortgage where the borrower is required to pay a penalty in the event of certain principal repayments.	Y = PPM N = Not PPM
Amort_Type	AMORTISATION TYPE	Fixed-rate mortgage or adjustable-rate mortgage.	FRM = Fixed-Rate Mortgage ARM = Adjustable-Rate Mortgage
Porperty_State	PROPERTY STATE	State where property is located.	Two-letter abbreviation
Property_Type	PROPERTY TYPE	Type of property.	CO = Condominium PU = Planned Unit Development (PUD) MH = Manufactured House SF = Single-Family Home CP = Cooperative Share NaN = Not available
ZIP_Code	POSTAL CODE	Three-digit ZIP code: last two digits are obscured (00) for privacy.	### NaN = Unknown
Loan_Purpose	LOAN PURPOSE	Purchase: mortgage for the purchase of a property. Cash-Out Refinance: loan amount not restricted to specific purposes. No Cash-Out Refinance: Loan amount is limited to pay off first mortgage, pay off junior liens secured by mortgage property, pay	P = Purchase C = Cash-Out Refinance N = No Cash-Out Refinance

		related closing costs, finance charges and quick items, disburse cash out to borrower (no more than 2% or $2000).	R = Refinance Not Specified NaN = Not Available
Loan_Term	ORIGINAL LOAN TERM	Number of scheduled monthly payments.	(Loan Maturity Date (YYYY-MM) - Loan First Payment Date (YYYY-MM) + 1)
Num_Borrowers	NUMBER OF BORROWERS	Number of borrower(s) who are obligated to repay the mortgage	2018Q1 and prior: 01 = 1 borrower 02 = > 1 borrower 2018Q2 and later: 01 = 1 borrower 02 = 2 borrowers … 10 = 10 borrowers NaN = Not Available
Sup_Conforming_Flag	SUPER CONFORMING FLAG	Mortgage exceeds conforming loan limits in terms of principal balance.	Y = Yes N = Not Super Conforming
Ref_Ind	RELIEF REFINANCE INDICATOR	Whether the loan is part of Freddie Mac's Relief Refinance Program.	Y = Yes N = Non-Relief Refinance Loan
Property_Val_Meth	PROPERTY VALUATION METHOD	Method used to the obtain property appraisal, if any.	AC = ACE Loans A = Full Appraisal O = Other Appraisal ACP = ACE + PDR R = GSE Targeted Refinance NaN = Not Available

Int_Only	INTEREST ONLY INDICATOR	Whether loan only requires interest payments.	Y = Yes N = No
Loan_Num	LOAN SEQUENCE NUMBER	Unique loan identifier.	PYYQnXXXXXXX P = Product (F or A) YYQn = Origination year and Quarter
Month	MONTHLY REPORTING PERIOD	Month of loan information contained in the record.	YYYY-MM
Curr_UPB	CURRENT ACTUAL UPB	Ending mortgage balance for the corresponding reference period. May include non-interest-bearing deferred amounts in the event of modification or forbearance. Set to zero if Zero_Bal_Code is populated.	(Interest Bearing UPB) + (Non-Interest Bearing UPB)
Delinquency_Status	CURRENT LOAN DELINQUENCY STATUS	Value corresponding to the number of days the borrower is in arrears, based on the date of the last instalment paid (DDLPI).	0 = Current or less than 30 days delinquent 1 = 30-59 days delinquent 2 = 60-89 days delinquent 3= 90 -119 days delinquent And so on… RA = REO acquisition NaN = Not Available
Loan_Age	LOAN AGE	Number of scheduled payments (months) from origination up to and including the current reporting period. For modified loans, from the first payment date after the modification. Deferred payments are not considered (no modification).	Monthly Reporting Period (YYYY-MM) - Loan First Payment Date (YYYY-MM) + 1 Or Monthly Reporting Period (YYYY-MM) - Modification First Payment Date (YYYY-MM) + 1

Months_to_Maturity	REMAINING MONTHS TO LEGAL MATURITY	Number of months remaining to mortgage maturity or modified maturity date.	Maturity Date (YYYY-MM) - Monthly Reporting Period (YYYY-MM) + 1 Or Modified Maturity Date (YYYY-MM) - Monthly Reporting Period (YYYY-MM) + 1
Modification_Flag	MODIFICATION FLAG	Flag indicating whether the mortgage has been modified in the current period or in a previous period.	Y = current period modification P = prior period modification N = Not modified
Zero_Bal_Code	ZERO BALANCE CODE	Code for the reason why the UPB of the loan was reduced to zero.	01 = Prepaid or Matured 02 = Third Party Sale (DEFAULT) 03 = Short Sale or Charge Off (DEFAULT) 06 = Repurchased 09 = REO Disposition (DEFAULT) 15 = Whole Loan Sales or Non-Performing Sale 16 = Reperforming Loan Securitizations or Sale 96 = Defect prior to other termination events or Removal
Curr_Int_Rate	CURRENT INTEREST RATE	Current interest rate on the mortgage note, taking into account any loan modifications.	

Payment_Deferral_Flag	PAYMENT DEFERRAL	Whether the loan was granted a deferral in the current or previous period.	Y = Current Period P = Prior Period NaN = Not Applicable or Not Available
Assistance_Status	BORROWER ASSISTANCE STATUS FLAG	Type of assistance plan the borrower is enrolled in that provides temporary mortgage payment relief (regardless of delinquency status).	F = Forbearance R = Repayment T = Trial Period O = Other Workout Plan NaN = No plan, Not Applicable or Not Available
Payment_Ratio	PAYMENT RATIO	Proportion of the original UPB that has already been repaid.	Percentage [0,1]
Balance_Per_Month	PRINCIPAL PAYMENT AMOUNT PER MONTH	Approximation of the principal obligation per month until maturity.	
Delinq_Status_1_Count	30 DPD COUNT	Number of times the borrower has been 30 days in arrears in the past 12 months.	Range 0-12
Delinq_Status_2_Count	60 DPD COUNT	Number of times the borrower has been 60 days in arrears in the past 12 months.	Range 0-12
Delinq_Status_3+_Count	90+ DPD COUNT	Number of times the borrower has been 90+ days in arrears in the past 12 months.	Range 0-12
Max_Delinq_Severity_yoy	MAXIMUM DELINQUENCY SEVERITY IN THE PAS 12 MONTHS	Highest level of delinquency (30 DPD, 60 DPD or 90+ DPD) recorded in the past 12 months.	
Num_Modifications_yoy	NUMBER OF MODIFICATIONS IN THE PAST 12 MONTHS	Number of loan modifications in the past 12 months.	

Unemp_Rate	UNEMPLOYMENT RATE	Unemployment rate at the county level.	
Median_Ann_Icome	MEDIAN ANNUAL INCOME	Median annual income at the state level.	
Infl_Rate	INFLATION RATE	U.S. inflation rate.	
HP_Index	HOUSE PRICE INDEX	House price index (FHFA) at the ZIP code level.	
US_30yr_FRM	CURRENT U.S. 30-YEAR FIXED RATE MORTGAGE AVERAGE	Current U.S. 30-year fixed-rate mortgage average (Freddie Mac).	
US_30yr_FRM_Orig	U.S. 30-YEAR FIXED RATE MORTGAGE AVERAGE AT ORIGINATION	U.S. 30-year fixed-rate mortgage average (Freddie Mac) at mortgage origination.	
HPI_Change_Orig	HPI PERCENTAGE CHANGE SINCE ORIGINATION	Percentage change in the house price index (FHFA) since mortgage origination.	
HPI_Change_yoy	HPI PERCENTAGE CHANGE IN THE PAST 12 MONTHS	Percentage change in the house price index (FHFA) in the past 12 months.	
Int_Rate_Diff	ORIGINATION INTEREST RATE AND NATIONAL MORTGAGE RATE DIFFERENCE	Difference between the loan interest rate at origination and the National Mortgage Rate (Freddie Mac) at origination.	
Curr_Int_Rate_Diff	CURRENT INTEREST RATE AND NATIONAL MORTGAGE RATE DIFFERENCE	Difference between the current loan interest rate and the current National Mortgage Rate (Freddie Mac).	

Appendix B: Termination Events

Termination Event	**Definition**
Maturity	The entire outstanding balance, together with any accrued interest, has been repaid by the borrower, thereby terminating the mortgage agreement.
Voluntary prepayment	The borrower decides to repay all or part of the mortgage before the scheduled maturity date.
Whole Loan Sale	Sale of the mortgage to another entity, resulting in the removal of the loan from the seller's books and the transfer of all related rights and obligations, such as the collection of payments, to the purchaser.
Short Sale	Sale of the property by the homeowner (borrower), with the consent of the lender, for less than the outstanding unpaid balance. The difference is still owed by the borrower.
Charge-off	Writing off a mortgage as a loss after the foreclosure process has been completed. Foreclosure occurs when the borrower stops repaying the debt (goes into default) and the lender attempts to recover the outstanding unpaid balance by selling the property at auction or through a direct transaction. If the sale price of the property does not cover the amount owed by the borrower, the lender must record a loss.
Third Party Sale	Sale of the property to a third party other than the borrower or lender, typically during the foreclosure process.
REO Disposition	Sale of a Real Estate Owned (REO) property. A property is called REO when the lender repossesses it after the borrower defaults and the property has not been sold during the foreclosure process. The lender will try to sell the property as quickly as possible, often at a discount, to cover at least part of the outstanding debt.

Reperforming Loan Securitisation	Pooling mortgages that were previously in default but are now making regular payments, and then issuing securities backed by these loans. This allows the lender to manage its risk and generate liquidity.

Appendix C: Sparse Generalized Additive Model

$$
\begin{aligned}
\text{score} = {} & -1.7025292645871624 \\
& -0.2507023347432949 \times 1_{\text{Occupacy\_Status=Second Home}} \\
& -0.3816688037282125 \times 1_{\text{Channel=Retail}} \\
& +0.11049490860211929 \times 1_{\text{Loan\_Purpose=No Cash-Out Refinance}} \\
& +1.086758118222201 \times 1_{\text{Assistance\_Status=Repayment}} \\
& -1.2028775781808012 \times 1_{\text{Assistance\_Status=Trial Period}} \\
& +0.4693001124221724 \times 1_{\text{Modification\_Flag=Prev. Period}} \\
& +0.4692235961968327 \times 1_{\text{Payment\_Deferral\_Flag=Yes}} \\
& +0.22335282961803038 \times 1_{\text{Payment\_Deferral\_Flag=Prev. Period}} \\
& +0.2182935774524996 \times 1_{\text{First\_Time\_Home\_Buyer\_Flag=Yes}} \\
& +0.19302676013673 \times 1_{\text{Prepay\_Penalty=Yes}} \\
& +0.17936958813228615 \times 1_{\text{Num\_Units=Missing}} \\
& +0.3956432443483549 \times 1_{\text{Credit\_Score}\leq 600} \\
& +0.3086993202849177 \times 1_{\text{Credit\_Score}\leq 700} \\
& +0.29216096167903544 \times 1_{\text{Credit\_Score}\leq 750} \\
& +0.49809197634570107 \times 1_{\text{Credit\_Score}\leq 800} \\
& -0.37755069621208487 \times 1_{\text{DTI}\leq 28.44} \\
& -0.18720486015421364 \times 1_{\text{LTV}\leq 90.66} \\
& +0.2304500383111024 \times 1_{\text{Median\_Ann\_Income}\leq 37707.8} \\
& +0.15458914455037503 \times 1_{\text{US\_30yr\_FRM}\leq 3.02} \\
& +0.22279868666722702 \times 1_{\text{Payment\_Ratio}\leq 0.04} \\
& +0.2118268842411819 \times 1_{\text{Payment\_Ratio}\leq 0.06} \\
& +0.21481725048277023 \times 1_{\text{Payment\_Ratio}\leq 0.1} \\
& +0.33966204443702125 \times 1_{\text{Payment\_Ratio}\leq 0.16} \\
& -1.978544333687718 \times 1_{\text{Delinquency\_Status}\leq 1.0} \\
& -1.0012730387945605 \times 1_{\text{Delinquency\_Status}\leq 2.0} \\
& +0.5284206912741195 \times 1_{\text{Num\_Borrowers}\leq 2.0} \\
& -0.39430784909536165 \times 1_{\text{Loan\_Age}\leq 13} \\
& +0.5364447679026803 \times 1_{\text{Loan\_Age}\leq 29} \\
& -1.7102192512215189 \times 1_{\text{Max\_Delinq\_Severity\_yoy}\leq 2}
\end{aligned}
$$

References

Angelino, E., Larus-Stone, N., Alabi, D., Margor, S., & Rudin, C. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*, 1-78.

Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable FInTech Lending. *Journal of Economics and Business*, 125-126.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 627-635.

Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting Loan Default in Europe with Machine Learning. *Journal of Financial Econometrics*, 569-596.

Bartlett, M. S. (2021). Consumer-lending discrimintaion in the FinTech Era. *Journal of Financial Economics*.

Bussman, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, 203-216.

Butaru, F., Chen, Q., Clark, B., Das, S., W. Lo, A., & Siddique, A. (2016). Risk and Risk Management in the Credit Card Industry. *Journal of Banking & Finance*, 218-239.

Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006). Getting the most out of ensemble selection. *Proceeding of the 6th IEEE international conference on data mining* , 828-833.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 321-357.

Chen, C., Kangcheng, L., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An Interpretable Model with Globally Consisten Explanations for Credit Risk. *NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy.* Montréal, Canada.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 785-794).

Croxson, K., Bracke, P., & Jung, C. (2019, May 31). Explaining why the computer say 'no'. *Insight: Opinion and Analysis hosted by the FCA*.

Dedieu, A., Hazimeh, H., & Mazumder, R. (2021). Learning Sparse Classifiers: Continuous and Mixed Integer Optimization Perspectives. *Journal of Machine Learning Research*, 1-47.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 37-54.

FHFA. (2024). *Fannie Mae and Freddie Mac Single-Family Guarantee Fees 2022.*

Florez-Lopez, R. (2010). Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society*, 486-501.

Freddie Mac. (2024). *Single-Family Loan-Level Dataset General User Guide.*

Freund, Y., & Schapire, R. E. (1997). A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 119-139.

FSB, F. S. (2017). *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications.*

Glennon, K. L. (2008). Development and Validation of Credit-Scoring Models.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *Neural Information Processing Systems*.

Hazimeh, H., Mazumder, R., & Nonet, T. (2022). L0Learn: A Scalable Package for Sparse Learning using L0 Regularization. *arXiv* , 2202.04820.

Ho, T. K. (1995). Random Decision Forests. *3rd International Conference on Document Analysis and Recognition*, (pp. 278-282). Montreal.

Hussein, H., & Rahul, M. (2020). Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms. *Operations Research*, 1517-1537.

Kennedy, K., Mac Name, B., Delany, S. .., O'Sullivan, M., & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications*, 1372-1380.

Khandani, A. E., Adlar, J. K., & Andrew, W. L. (2010). Consumer Credit Risk Models via Machine-Learning Algorithms. *Journal of Banking & Finance 34*, 2767-2787.

Lessmann, S., Baesens, B., Seow, H.-V., & Lyn C., T. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 124-136.

Liu, J., Zhong, C., Seltzer, M., & Rudin, C. (2022). Fast Sparse Classification for Generalized Linear and Additive Models. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS).* Valencia, Spain.

Liu, J., Zhong, C., Seltzer, M., & Rudin, C. (2022). Fast Sparse Classification for Generalized Linear and Additive Models. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, (pp. 9304-9333). Valencia, Spain.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Misheva, B. H., Hirsa, A., Osterrieder, J., Kulkarni, O., & Lin, F. S. (2021). Explainable AI in Credit Risk Management.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometrics Approach. *Journal of Economic Perspectives*, 87-106.

Murdoch, S. K.-A. (2019). Definitions, methods, and applications in interpretable machine learning. *PNAS*.

Ojha, V., & JeongHoe, L. (2021). Default analysis in mortgage risk with conventional and deep machine learning focusing on 2008-2009. *Digital Finance*, 249-271.

Rahmani, R., Parola, M., & Cimino, M. G. (2024). A machine learning workflow to address credit default prediction.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge and Data Mining*.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell.*, 206-2015.

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*.

Sharan, V., Tai, K. S., Bailis, P., & Valiant, G. (2017). There and Back Again: A General Approach to Learning Sparse Models.

Siddiqi, N. (2017). *Intelligent Risk Scoring* (2nd Edition ed.). John Wiley & Sons.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2020). Deep Learning for Mortgage Risk. *Journal of Financial Econometrics*, 313-368.

Tibshirani, R., & Hastie, T. (1986). Generalized Additive Models. *Statistical Sciences*, 297-318.